Online video game store

Problem Description

The objective of this project is to mine association rules between the video games purchased by customers from an online video game store. This is a classic problem of Association Rule Learning, which is a technique used to find interesting relationships between variables in large datasets.

Dataset Description

The dataset used for this project is from Kaggle and contains transactional data from an online video game store. The dataset includes information on the name of the game, the platform it was released on, the year of release, the genre, the publisher, the region of sale, and the number of copies sold in millions. The dataset contains data from 1980 to 2016.

Background Information

Association Rule Learning is a type of unsupervised learning where the goal is to find interesting relationships between variables in large datasets. The most common application of Association Rule Learning is in market basket analysis, where the goal is to identify which items are frequently purchased together by customers. This information can be used by retailers to optimize their product placements and marketing strategies.

In this project, we will be using the Apriori algorithm, which is a classic algorithm used for mining association rules in large datasets. The algorithm works by generating candidate itemsets and then pruning them based on a minimum support threshold. The remaining itemsets are then used to generate the association rules.

The Apriori algorithm is based on the following principles:

- **Support**: This is the proportion of transactions in which an itemset appears. The higher the support, the more frequent the itemset.

- **Confidence**: This is the proportion of transactions containing the antecedent that also contain the consequent. The higher the confidence, the more likely the consequent is to be purchased when the antecedent is purchased.
- **Lift**: This is the ratio of observed support to expected support. It measures the strength of the association rule between the antecedent and the consequent. A lift of 1 indicates that the antecedent and consequent are independent, while a lift greater than 1 indicates a positive association between the two.

  The goal of this project is to use the Apriori algorithm to find interesting association rules between video games purchased by customers from the online video game store.

Possible Framework :

1. **Data Preprocessing:**
- Load the dataset and inspect its properties.
- Drop any irrelevant or redundant columns.
- Handle missing values and duplicates.
- Encode categorical features into numerical values.
- Transform data into a transaction format suitable for association rule mining.
2. **Association Rule Mining:**
- Identify frequent itemsets using a suitable algorithm (e.g. Apriori, FP-Growth).
- Generate association rules from the frequent itemsets with a minimum support and confidence threshold.
- Evaluate the quality of the association rules using metrics such as lift, conviction, and interest.
- Select and interpret the most interesting and actionable association rules.
3. **Visualization:**
- Visualize the frequent itemsets and association rules in a clear and understandable way.
- Use appropriate visualization techniques (e.g. scatterplots, heatmaps, network graphs) to highlight interesting patterns and relationships.
4. **Interpretation and Recommendation:**
- Interpret the patterns and relationships found in the association rules.
- Identify actionable insights and recommendations for the online video game store.
- Use the association rules to suggest relevant games to customers based on their past purchases.
5. **Deployment:**
- Build an API or web interface for the association rule model.
- Deploy the model to a suitable cloud platform.
- Test the model on new data and evaluate its performance.
- Monitor the model for changes in data or patterns and update as necessary.

Code Explanation :

Here is the simple explanation for the code which is provided in the code.py file.

Data Preprocessing

In this section, we load the dataset and perform necessary preprocessing steps such as handling missing values, dropping irrelevant columns, and converting data types. We have also used label encoding to convert categorical data into numerical values.

Association Rule Mining

In this section, we use the Apriori algorithm for Association Rule Mining. First, we perform one-hot encoding to transform the data into a format that can be used by the algorithm. Next, we use the apriori function to extract frequent itemsets and generate association rules based on a minimum support and confidence threshold. We also print the rules and their corresponding support, confidence, and lift scores.

Visualizing Association Rules

In this section, we use the seaborn library to create a heatmap of the support, confidence, and lift scores of the association rules. The heatmap allows us to quickly identify which rules have the highest scores and are thus the most interesting.

We have used the Apriori algorithm for association rule mining. It is a popular algorithm for mining frequent itemsets and generating association rules. The algorithm works by first identifying all frequent itemsets in the dataset and then using those itemsets to generate association rules.

To run this code, you will need to have Python installed along with the following libraries: pandas, numpy, sklearn, mlxtend, seaborn. You can install these libraries using pip. The dataset used for this project can be downloaded from the Kaggle website.

Motivation: By running this code, you will be able to extract association rules from a transactional dataset, which can be used to identify patterns and relationships between different items. This can be useful in a variety of applications, including market basket analysis and customer behavior analysis.

Future Work:

1. **Data Preprocessing:** In the future, more advanced techniques could be used for data preprocessing, such as data imputation or outlier detection to better handle missing values or noisy data.
2. **Hyperparameter Tuning:** The performance of the association rule mining model can be improved by tuning the hyperparameters used during the model training. This can be done by using a grid search or random search approach to identify the best set of hyperparameters.
3. **Feature Engineering:** Additional features can be extracted from the transactional data to improve the performance of the model. For example, one could use the time of day or day of the week to better capture patterns in customer behavior.
4. **Integration with Other Data Sources:** Incorporating data from other sources, such as demographic data or social media data, could provide additional insights into customer behavior and help identify new patterns and trends.
5. **Real-Time Recommendations:** The current model works on historical transactional data, but it can be adapted to provide real-time recommendations to customers as they are browsing the online store. This can be done by integrating the model with the online store's recommendation engine.

Step-by-Step Guide:

1. **Data Cleaning and Preprocessing:** As a first step, we need to clean and preprocess the transactional data to remove duplicates and handle any missing or noisy data. This can be done using pandas and other data preprocessing libraries.
2. **Association Rule Mining:** Once the data is preprocessed, we can use the Apriori algorithm to mine association rules between different games purchased by customers. This can be done using the mlxtend library in Python.
3. **Evaluation and Hyperparameter Tuning:** After mining the association rules, we need to evaluate the performance of the model and tune the hyperparameters to improve its accuracy. We can use metrics such as support, confidence, and lift to evaluate the quality of the association rules.
4. **Feature Engineering:** Additional features can be extracted from the transactional data to improve the performance of the model. For example, one could use the time of day or day of the week to better capture patterns in customer behavior.
5. **Integration with Other Data Sources:** We can incorporate data from other sources, such as demographic data or social media data, to provide additional insights into customer behavior and help identify new patterns and trends.
6. **Real-Time Recommendations:** Once the model is trained and evaluated, it can be integrated with the online store's recommendation engine to provide real-time

recommendations to customers as they browse the store. This can help improve customer engagement and increase sales.

Exercise :

Try to answers the following questions by yourself to check your understanding for this project. If stuck, detailed answers for the questions are also provided.

What is association rule learning and how does it work in the context of this project?
Answer: Association rule learning is a method of discovering interesting relationships between variables in large datasets. In the context of this project, it involves discovering patterns in the video game purchase history of customers. This is done by calculating support, confidence and lift values for all possible combinations of games and selecting those with the highest scores.

How would you evaluate the performance of the association rule mining model?
Answer: There are several measures that can be used to evaluate the performance of an association rule mining model, such as support, confidence and lift values. Additionally, one can also use metrics like accuracy and F1-score to evaluate the model's performance.

Can you explain the difference between support, confidence and lift values?

Answer: Support refers to the proportion of transactions in the dataset that contain a specific itemset. Confidence is the proportion of transactions that contain both an antecedent and a consequent itemset, out of all transactions that contain the antecedent. Lift measures the strength of association between an antecedent and consequent itemset, while taking into account the support of both itemsets.

Can you describe any potential issues or limitations of using association rule mining for this project?

Answer: One potential limitation of association rule mining is that it may not capture all relevant patterns in the data, as it only considers pairwise relationships between items. Additionally, association rule mining can be computationally expensive for large datasets and may require significant preprocessing to remove noisy or irrelevant data.

How can you use association rule mining to improve sales and customer satisfaction in the video game store?

Answer: One way to use association rule mining to improve sales and customer satisfaction is to identify which games are commonly purchased together, and then create targeted promotions or bundles that include those games. Additionally, identifying frequent itemsets can help identify popular game genres or themes, which can inform decisions about which new games to stock in the store.

Concept Explanation :

Imagine you go grocery shopping, and you have a list of items that you need to buy. But what if you could predict what other items you might want to purchase based on what's already in your cart? That's essentially what Apriori algorithm does in a nutshell.

The algorithm works by looking at frequent itemsets, which are combinations of items that appear frequently together in the data. It then uses these frequent itemsets to generate rules that can be used to make predictions about what items are likely to be purchased together.

To achieve this, the algorithm first scans the dataset to find all frequent itemsets, which are itemsets that appear above a certain threshold frequency (called the support threshold). It then uses these frequent itemsets to generate association rules, which are rules of the form "if you buy item A, you are likely to buy item B".

These association rules are generated based on a measure called confidence, which measures the likelihood that item B is purchased given that item A is purchased. The algorithm then selects the rules with the highest confidence values to recommend to the user.

Overall, the Apriori algorithm is a powerful tool for discovering hidden relationships between items in a dataset, and can be used to make predictions about what items are likely to be purchased together. It's commonly used in market basket analysis, recommendation systems, and other similar applications.

I hope this explanation was helpful and made you excited to learn more about Apriori algorithm!