AIC-201 Supervised and Unsupervised Machine Learning
Assessment Activity AIC-201:00040
Unsupervised Learning

Name: (Win) Thanawin Pattanaphol Student ID: 01324096

Part 1

Difference: K-means clustering vs Hierarchical clustering:

K-means Clustering

- A centroid-based clustering algorithm that divides the data into a specific number of clusters. It decreases the variance within clusters by continuously updating centroids – requires a pre-defined set amount of clusters.

Hierarchical Clustering

- Tree-based clustering which creates a hierarchy of clusters where it can either be a bottom-up approach or top-down approach – to calculate the number of clusters, it can be found by cutting the tree diagram (dendrogram) at a specific level.

K-means clustering would be a better method for customer segment due to the following reasons:

- **Scalable**
  K-means is more efficient with large datasets as it scales better compared to hierarchical clustering.

- **Clarity**
  K-means provides clear cluster centroids, thus, easier to read and understand customer segments in terms of shopping behavior.

- **Real-World Practicality**
  As retails often need to segment customers based on predefined criteria; K-means is a better method as it allows for flexible iteration over different *k* values to determine the most important number of clusters.

- **Limitations of Hierarchical Clustering:** It is computationally intensive and may not offer an easily readable or practical partitioning of customers.

Recommended Data Preprocessing Pipeline

1. **Filling Missing Values**
   Any missing values in any of then features can be filled in using the media as it is resistant to outliers.
2. **Working with Outliers**

   Use methods such as Z-score or the IQR (Interquartile Range) to detect outliers and cap or get rid of any of the outlier values.
3. **Scaling Features**

   Standardize the data using z-score normalization or via min-max scale (transforming values in the range of 0-1)

Pseudocode for K-means Clustering

```
centroids = random_k_points(data, k)

while not converged:
    for each data_point in data:
        assign_to_nearest_centroid(data_point, centroids)

    for each centroid in centroids:
        new_centroid = mean_of_assigned_points(centroid)
        update_centroid(centroid, new_centroid)

    if centroids_converged():
        break
```

Strategies for Engaging Customers from Each Cluster:

- **Cluster 1 (High frequency, low monetary, and recent transactions):**

  Provide personalized promotions or discounts on high-value items or loyalty programs which can lead to an increasing spending per transaction.
- **Cluster 2 (Low frequency, high monetary, older transactions):**
  - Create re-engagement communication with exclusive offers on specific genres of products and provide incentives such as free shipping costs or discounts to encourage repeat purchases.
- **Cluster 3 (Medium frequency, medium monetary, and moderately recent transactions):**

  As these customers are moderately engaged and have a moderate shopping behavior, it is recommended to create personalized incentives to increase both frequency and monetary value, such as, recommendations based on past purchases or other types of marketing campaigns.

Part 2

Question 3:

Support: The proportion of transactions in the dataset that contain a particular itemset. It is a measure of how frequently an item or itemset occurs in the dataset.

Formula:

$$Support(A \rightarrow B) = \frac{\text{Number of transactions containing both A \& B}}{\text{Total number of transactions}}$$

Confidence: The probability that item Y is purchased when item X is purchased. It measures the likelihood of finding item Y in transactions where there is item X.

Formula:

$$Confidence(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support(A)}$$

Lift: The ratio of the observed support to the expected support if X and Y were independent.

$$Lift(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support(A) \times Support(B)}$$

A lift greater than 1 indicates that the items appear together more often than would be expected by chance – a positive correlation between A & B.

Example Dataset:

Transaction ID	Purchased items
T1	Milk, Bread, Butter
T2	Milk, Bread
T3	Milk, Butter
T4	Bread, Butter
T5	Milk, Bread, Butter

Deriving Rule

Let rule be: {Milk} ⇒ {Bread}

Support: The number of transactions where both Milk and Bread are purchased, divided by the total number of transactions

$$Support(Milk \rightarrow Bread) = \frac{3}{5} = 0.6$$

Confidence: Support of Milk, Bread divided by the support of Milk.

$$Confidence(Milk \rightarrow Bread) = \frac{0.6}{0.8} = 0.75$$

Lift: Support of the rule divided by the product of the supports of Milk and Bread

$$Lift(Milk \rightarrow Bread) = \frac{0.6}{0.8 \times 0.8} = \frac{0.6}{0.64} = 0.9375$$

Question 4: Applying Association Rules (30 Marks)

1. **Using Apriori Algorithm to Generate Association Rules:**
    - Identify frequent itemsets that meet a minimum support threshold. This can be done by scanning the dataset for items that appear together frequently.
    - Generate candidate itemsets of length *k+1* by combining frequent itemsets of length *k*. Continue until no more frequent itemsets can be generated.
    - For each frequent itemset, generate association rules and calculate confidence. Only keep rules that meet a predefined confidence threshold.
    - **Threshold Selection:**
        - **Support:** Typically set to around 0.05 or 5%, meaning the rule must appear in at least 5% of the transactions.
        - **Confidence:** Typically set to around 0.5 or 50%, meaning the rule must hold in at least 50% of the transactions where the antecedent is present.
2. **Interpretation of Rule {Milk, Bread} → {Butter}:**
    - **Support:** 0.05 means this combination appears in 5% of all transactions.
    - **Confidence:** 0.6 means 60% of the transactions with Milk and Bread also include Butter.
    - **Lift:** 1.5 indicates that Milk and Bread are 1.5 times more likely to be bought together with Butter than by chance.

    Recommendation:

    - The grocery store can use this rule to create targeted promotions. For example, offering a discount on Butter when customers purchase Milk and Bread together could increase sales of Butter.
    - The store can display Butter near Milk and Bread to encourage customers to purchase all three items together, increasing the association to increase total transaction value.