Data Mining, Assignment 2

Shihao Liang, UID: 3O36196673

November 2023

1 Question 1: Classification

This report details the data mining process undertaken to develop a classification model for predicting the acceptance of in-vehicle coupon recommendations. The process encompasses various stages, including data preprocessing, attribute selection, parameter tuning, construction of training and testing sets, model building, and evaluation.

1.1 a

1.1.1 Data Preprocessing

Initial Inspection. The dataset comprised several attributes related to the circumstances under which coupons were offered to individuals in vehicles.

Attribute Removal. As per the requirement, the attributes "car", "Bar", "CoffeeHouse", "CarryAway", "RestaurantLessThan20", and "Restaurant20To50" were removed. This step was essential to focus on more relevant predictors for coupon acceptance.

Nominal Attribute Treatment All remaining attributes were treated as nominal. This treatment is suitable for models handling categorical data, such as decision trees.

1.1.2 Attribute Selection

One-Hot Encoding: After treating all attributes as nominal, we applied one-hot encoding. This step transformed categorical variables into a form that could be provided to machine learning algorithms to improve prediction accuracy.

1.1.3 Training and Testing Data Construction

Data Splitting: The dataset was split into training and testing sets using an 80-20 split. This standard practice in machine learning ensures that the model is tested on unseen data, providing a realistic assessment of its performance.

1.1.4 Model Building

Decision Tree Classifier A decision tree classifier was chosen for its suitability in handling categorical data and ease of interpretation.

Parameter Tuning Parameter tuning was performed using GridSearchCV, focusing on 'max_depth' and 'min_samples_split'. This step is crucial for optimizing the model's performance.

1.1.5 Model Evaluation

Accuracy The final model achieved an accuracy of approximately 64.01% on the test set.

Precision, Recall, and F1-Score The model showed a higher precision, recall, and F1-score for predicting coupon acceptance (class 1) compared to rejection (class 0).

Confusion Matrix Analysis The confusion matrix provided insights into the true positives, true negatives, false positives, and false negatives.

1.1.6 Conclusion and Recommendations

The decision tree classifier demonstrated moderate predictive capability. There is room for improvement in accuracy, which might be achieved through more sophisticated models or additional feature engineering.

Future work could explore alternative classifiers like Random Forests or Gradient Boosting.

Ongoing model refinement, including feature engineering and exploration of different classification algorithms, is recommended for enhanced performance.

1.1.7 Final Remarks

This report outlined the systematic approach taken to build and evaluate a classification model for the in-vehicle coupon recommendation dataset. Through meticulous preprocessing, careful attribute selection, and model optimization, a functional model was developed, with insights into potential areas for future improvements.

1.2 b

The prediction of the given three records is: **0, 1, 0**

2 Question 2: Association Analysis

2.1 a

Based on the 5-itemset Brioche, Beef, Cheese, Lettuce, Mushroom being frequent, the following 4-itemsets must also be frequent:

- {Brioche, Beef, Cheese, Lettuce}
- {Brioche, Beef, Cheese, Mushroom}
- {Brioche, Beef, Lettuce, Mushroom}
- {Brioche, Cheese, Lettuce, Mushroom}
- {Beef, Cheese, Lettuce, Mushroom}

These 4-itemsets are derived by considering all possible combinations of 4 items from the original 5-itemset.

2.2 b

Possible 4-itemsets and their support counts:

4-itemset	**Support Count**
{Bagel, Beef, Cheese, Lettuce}	0
{Bagel, Beef, Cheese, Mushroom}	5
{Bagel, Beef, Cheese, Pickle}	5
{Bagel, Beef, Lettuce, Mushroom}	2
{Bagel, Beef, Lettuce, Pickle}	2
{Bagel, Beef, Mushroom, Pickle}	7
{Bagel, Chicken, Cheese, Lettuce}	3
{Bagel, Chicken, Cheese, Mushroom}	3
{Bagel, Chicken, Cheese, Pickle}	3
{Bagel, Chicken, Lettuce, Mushroom}	9
{Bagel, Chicken, Lettuce, Pickle}	3
{Bagel, Chicken, Mushroom, Pickle}	3
{Bagel, Cheese, Lettuce, Mushroom}	3
{Bagel, Cheese, Lettuce, Pickle}	3
{Bagel, Cheese, Mushroom, Pickle}	8
{Bagel, Lettuce, Mushroom, Pickle}	5
{Brioche, Beef, Cheese, Pickle}	14
{Brioche, Beef, Lettuce, Pickle}	14
{Brioche, Beef, Mushroom, Pickle}	8
{Brioche, Chicken, Cheese, Lettuce}	8
{Brioche, Chicken, Cheese, Mushroom}	0
{Brioche, Chicken, Cheese, Pickle}	0
{Brioche, Chicken, Lettuce, Mushroom}	0
{Brioche, Chicken, Lettuce, Pickle}	0
{Brioche, Chicken, Mushroom, Pickle}	0
{Brioche, Cheese, Lettuce, Pickle}	14
{Brioche, Cheese, Mushroom, Pickle}	8
{Brioche, Lettuce, Mushroom, Pickle}	8
{Beef, Cheese, Lettuce, Pickle}	14
{Beef, Cheese, Mushroom, Pickle}	13
{Beef, Lettuce, Mushroom, Pickle}	10
{Chicken, Cheese, Lettuce, Mushroom}	3
{Chicken, Cheese, Lettuce, Pickle}	3
{Chicken, Cheese, Mushroom, Pickle}	3
{Chicken, Lettuce, Mushroom, Pickle}	3
{Cheese, Lettuce, Mushroom, Pickle}	11

2.3 c

The following table lists some itemsets and their respective support counts:

Itemset	**Support Count**
{Brioche, Beef, Cheese, Lettuce}	32
{Brioche, Beef, Cheese, Mushroom}	18
{Brioche, Beef, Cheese, Pickle}	14
{Brioche, Beef, Lettuce, Mushroom}	18
{Brioche, Beef, Lettuce, Pickle}	14
{Brioche, Cheese, Lettuce, Mushroom}	18
{Brioche, Cheese, Lettuce, Pickle}	14
{Beef, Cheese, Lettuce, Mushroom}	18
{Beef, Cheese, Lettuce, Pickle}	14
{Beef, Cheese, Mushroom, Pickle}	13
{Beef, Lettuce, Mushroom, Pickle}	10
{Cheese, Lettuce, Mushroom, Pickle}	11

2.4 d

According to the L4 obtained from part c), the C5 can be generated through the apriori-gen algorithm. The C5 includes:

- {Brioche, Beef, Cheese, Lettuce, Mushroom}
- {Brioche, Beef, Cheese, Lettuce, Pickle}
- {Brioche, Beef, Cheese, Pickle, Mushroom}
- {Brioche, Beef, Mushroom, Lettuce, Pickle}
- {Brioche, Mushroom, Cheese, Lettuce, Pickle}
- {Beef, Cheese, Lettuce, Pickle, Mushroom}

The support counts for each itemset in C5 are as follows:

Itemset	**Support Count**
{Brioche, Beef, Cheese, Lettuce, Mushroom}	18
{Brioche, Beef, Cheese, Lettuce, Pickle}	14
{Brioche, Beef, Cheese, Pickle, Mushroom}	8
{Brioche, Beef, Mushroom, Lettuce, Pickle}	8
{Brioche, Mushroom, Cheese, Lettuce, Pickle}	8
{Beef, Cheese, Lettuce, Pickle, Mushroom}	8

Therefore, the L5 includes {Brioche, Beef, Cheese, Lettuce, Mushroom} and {Brioche, Beef, Cheese, Lettuce, Pickle}, as their support counts are equal to or greater than 10.

3 Question 3: Cluster Analysis

3.1 Initialization

Commence with a singular cluster encompassing all eight points.

3.2 Bisecting

In the initial state, where only one cluster exists, the selection of a cluster for division is unnecessary.

Bisecting the Cluster using K(2)-Means The process of bisecting the cluster employs K(2)-means algorithm. The selection of initial centroids is based upon the Manhattan distances between pairs of objects within the cluster. The objective is to identify two objects that exhibit significant spatial separation. The Manhattan Distances (MD) between each pair of objects are listed as follows:

$$\begin{aligned}MD(O_1,O_2)=5, MD(O_1,O_3)=3, MD(O_1,O_4)=7, MD(O_1,O_5)=6, MD(O_1,O_6)=6,\\ MD(O_1,O_7)=8, MD(O_1,O_8)=8, MD(O_2,O_3)=8, MD(O_2,O_4)=4, MD(O_2,O_5)=11,\\ MD(O_2,O_6)=9, MD(O_2,O_7)=7, MD(O_2,O_8)=9, MD(O_3,O_4)=4, MD(O_3,O_5)=3,\\ MD(O_3,O_6)=3, MD(O_3,O_7)=5, MD(O_3,O_8)=5, MD(O_4,O_5)=7, MD(O_4,O_6)=5,\\ MD(O_4,O_7)=3, MD(O_4,O_8)=5, MD(O_5,O_6)=4, MD(O_5,O_7)=6, MD(O_5,O_8)=6,\\ MD(O_6,O_7)=2, MD(O_6,O_8)=2, MD(O_7,O_8)=2.\end{aligned}$$

Given that $MD(O_2,O_5)=11$ is the maximum distance observed, O_2 and O_5 are selected as the initial centroids for the K(2)-means process.

K(2)-means. Excluding points 2 and 3, the distances of all other points relative to the reference points 2 and 5 are compared, facilitating the categorization of these points into two distinct clusters.

The Manhattan Distances are defined as follows:

$$\begin{aligned}&MD(O_2,O_1)=5, && MD(O_2,O_3)=9, && MD(O_2,O_4)=4,\\ &MD(O_2,O_6)=9, && MD(O_2,O_7)=7, && MD(O_2,O_8)=9,\\ &MD(O_5,O_1)=6, && MD(O_5,O_3)=3, && MD(O_5,O_4)=7,\\ &MD(O_5,O_6)=4, && MD(O_5,O_7)=6, && MD(O_5,O_8)=6.\end{aligned}$$

Consequently, the points are divided into two clusters: $\{O_1,O_2,O_4\}$ and $\{O_3,O_5,O_6,O_7,O_8\}$.

For cluster $\{O_1, O_2, O_4\}$, the new centroid, denoted as centroid1, is positioned at coordinates $(4, 7)$. Similarly, for cluster $\{O_3, O_5, O_6, O_7, O_8\}$, the new centroid, denoted as centroid2, is located at $(7.6, 4.8)$.

The distances of each point to the two new centroids are computed, leading to a reevaluation of the cluster divisions.

$$\begin{aligned}
&MD(O_1, \text{new centroid1}) = 4, && MD(O_1, \text{new centroid2}) = 5.4\\
&MD(O_2, \text{new centroid1}) = 3, && MD(O_2, \text{new centroid2}) = 8.8\\
&MD(O_3, \text{new centroid1}) = 5, && MD(O_3, \text{new centroid2}) = 2.4\\
&MD(O_4, \text{new centroid1}) = 3, && MD(O_4, \text{new centroid2}) = 4.8\\
&MD(O_5, \text{new centroid1}) = 8, && MD(O_5, \text{new centroid2}) = 3.4\\
&MD(O_6, \text{new centroid1}) = 6, && MD(O_6, \text{new centroid2}) = 0.6\\
&MD(O_7, \text{new centroid1}) = 4, && MD(O_7, \text{new centroid2}) = 2.6\\
&MD(O_8, \text{new centroid1}) = 6, && MD(O_8, \text{new centroid2}) = 2.6
\end{aligned}$$

This analysis results in two clusters: $\{O_1, O_2, O_4\}$ and $\{O_3, O_5, O_6, O_7, O_8\}$, which remain identical to the previously established clusters. Thus, there is no requirement to recompute the centroids, and the K-2 means process is concluded.

To further refine the clusters, the Sum of Squared Errors (SSE) for each cluster is calculated, and the cluster with the higher SSE value will be chosen for splitting.

The SSE for cluster $\{O_1, O_2, O_4\}$ is calculated as follows:

$$SSE = \left[(|3-4| + |4-7|)^2\right] + \left[(|3-4| + |9-7|)^2\right] + \left[(|6-4| + |8-11|)^2\right] = 34$$

For cluster $\{O_3, O_5, O_6, O_7, O_8\}$, the SSE is:

$$\begin{aligned}
SSE = &\left[(|6-7.6| + |4-4.8|)^2\right] + \left[(|7-7.6| + |2-4.8|)^2\right]\\
&+ \left[(|8-7.6| + |5-4.8|)^2\right] + \left[(|8-7.6| + |7-4.8|)^2\right]\\
&+ \left[(|9-7.6| + |6-4.8|)^2\right] = 31.2
\end{aligned}$$

In conclusion, after recalculating the distances of each point to the respective centroids and reassessing the cluster divisions, the original clusters $\{O_1, O_2, O_4\}$ and $\{O_3, O_5, O_6, O_7, O_8\}$ remain unchanged. Therefore, it is not necessary to recompute the centroids. The process of K-2 means clustering is thus deemed complete. Additionally, the Sum of Squared Errors (SSE) for each cluster is computed, with the intention of identifying and splitting the cluster exhibiting the highest SSE.

3.3 Bisecting $\{O_1, O_2, O_4\}$

The Sum of Squared Errors (SSE) for cluster $\{O_1, O_2, O_4\}$ is observed to be greater than that of $\{O_3, O_5, O_6, O_7, O_8\}$. Therefore, the cluster $\{O_1, O_2, O_4\}$ is selected for subdivision.

1. *Selection of Initial Centroids*: The initial centroids are chosen based on their substantial separation. The Manhattan Distances are given by:

$$\begin{aligned}
MD(O_1, O_2) &= 5,\\
MD(O_1, O_4) &= 7,\\
MD(O_2, O_4) &= 4.
\end{aligned}$$

2. *Initial Cluster Division*: The points in $\{O_1, O_2, O_4\}$ are divided into two clusters based on their distances:

$$\begin{aligned}
MD(O_1, O_2) &= 5,\\
MD(O_2, O_4) &= 4.
\end{aligned}$$

This results in two clusters: $\{O_2, O_4\}$ and $\{O_1\}$.

3. *Recomputing the Centroids*:
    - For $\{O_1\}$, the new centroid (centroid1) is calculated as (3, 4).
    - For $\{O_2, O_4\}$, the new centroid (centroid2) is determined to be (4.5, 8.5).
4. *Re-dividing the Clusters*: The clusters are re-evaluated based on their distances to the new centroids:

$$\begin{aligned} MD(O_1, \text{new centroid1}) &= 0, & MD(O_1, \text{new centroid2}) &= 6, \\ MD(O_2, \text{new centroid1}) &= 5, & MD(O_2, \text{new centroid2}) &= 2, \\ MD(O_4, \text{new centroid1}) &= 7, & MD(O_4, \text{new centroid2}) &= 2. \end{aligned}$$

    The division remains as $\{O_1\}$ and $\{O_2, O_4\}$, concluding the K-2 means process.

3.4 Final Clustering

Ultimately, the points are divided into three clusters: $\{O_1\}$, $\{O_2, O_4\}$, $\{O_3, O_5 O_6, O_7, O_8\}$.