CPS 844: Data Mining

Assignment 2: Association analysis and Clustering analysis

Professor Elodie Lugez

Omar Syed - 500809837

A S M Rubayet Ahmed - 500962603

Background:

For our association analysis, we used a grocery store dataset from kaggle. Association rule learning in data mining is a rule based method for machine learning for relational discovery in datasets. Its purpose is to apply machine learning to distinguish strong rules present within the data set in an attempt to explain how and why the items in a data set are connected. The data set used for association analysis contained an unprocessed set of data with each data point having a small list of grocery items. The data set we used for clustering analysis is the iris dataset from the UCI database. Clustering Analysis is a form of a machine learning task that is identified as unsupervised. Clustering analysis works with identifying groupings in a data set to separate key functions. The data set used for our clustering analysis contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The classes are Iris Setosa, Iris Versicolour and Iris Virginica. The attributes are sepal length in cm, sepal width in cm, petal length in cm, petal width in cm and class: We chose this dataset as it is one of the most famous datasets in the pattern recognition domain, and is very suitable for clustering analysis. The dataset was created by R.A. Fisher.

Method:

For our association analysis, we began with a data filled with strings, since the majority of machine learning algorithms operate with the use of numerical data sets, as they function based off of mathematical computations, we had to begin by processing our data set to create a numerical data set. We did this by separating the list of grocery items into each cell, making them their own attribute. We then proceeded to indicate by binary values, 1 and 0, when a specific item in that attribute column recurred as the dataset went down the list.

```
[['MILK', 'BREAD', 'BISCUIT'], ['BREAD', 'MILK', 'BISCUIT', 'CORNFLAKES'], ['BREAD', 'TEA',
'BOURNVITA'], ['JAM', 'MAGGI', 'BREAD', 'MILK'], ['MAGGI', 'TEA', 'BISCUIT'], ['BREAD', 'TEA',
'BOURNVITA'], ['MAGGI', 'TEA', 'CORNFLAKES'], ['MAGGI', 'BREAD', 'TEA', 'BISCUIT'], ['JAM',
'MAGGI', 'BREAD', 'TEA'], ['BREAD', 'MILK'], ['COFFEE', 'COCK', 'BISCUIT', 'CORNFLAKES'],
['COFFEE', 'COCK', 'BISCUIT', 'CORNFLAKES'], ['COFFEE', 'SUGER', 'BOURNVITA'], ['BREAD', 'COFFEE',
'COCK'], ['BREAD', 'SUGER', 'BISCUIT'], ['COFFEE', 'SUGER', 'CORNFLAKES'], ['BREAD', 'SUGER',
'BOURNVITA'], ['BREAD', 'COFFEE', 'SUGER'], ['BREAD', 'COFFEE', 'SUGER'], ['TEA', 'MILK', 'COFFEE',
```

	BISCUIT	BOURNVITA	BREAD	COCK	COFFEE	...	JAM	MAGGI	MILK	SUGER	TEA
0	1	0	1	0	0	...	0	0	1	0	0
1	1	0	1	0	0	...	0	0	1	0	0
2	0	1	1	0	0	...	0	0	0	0	1
3	0	0	1	0	0	...	1	1	1	0	0
4	1	0	0	0	0	...	0	1	0	0	1
5	0	1	1	0	0	...	0	0	0	0	1
6	0	0	0	0	0	...	0	1	0	0	1
7	1	0	1	0	0	...	0	1	0	0	1
8	0	0	1	0	0	...	1	1	0	0	1
9	0	0	1	0	0	...	0	0	1	0	0
10	1	0	0	1	1	...	0	0	0	0	0

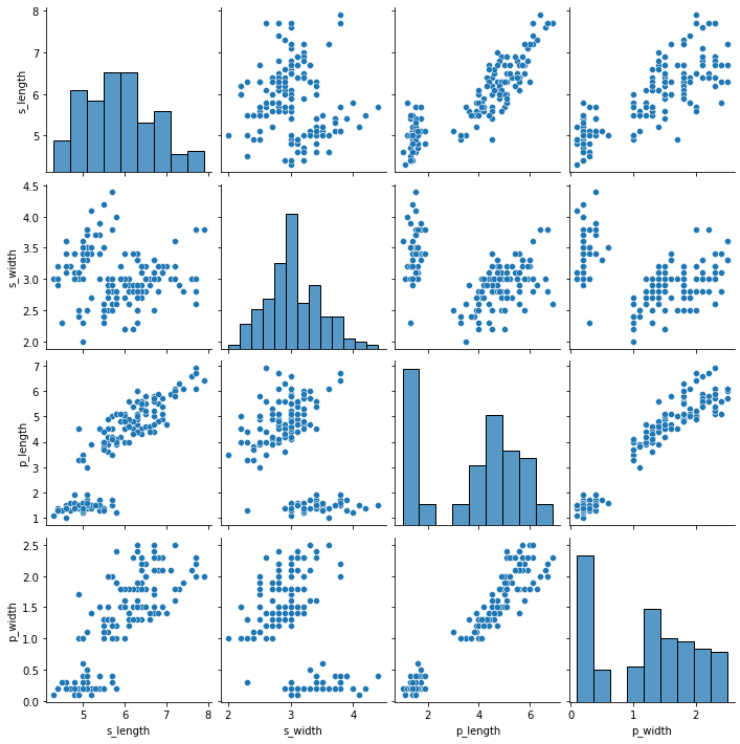
After applying binary classification to the set of items on the grocery lists, the next part required an association rule application. An antecedent, if rule, which operates in relation to the previous item. This antecedent if rule, operates with the calculation of probability, like such ex. If thing A occurs, there is a _% chance of thing b occurring. This ‘if’ relation helps find the probability of connections between combinations of data points. Similarly the rule of consequent is applied, which is referred to as the ‘then’ aspect of the whole correlation based rule (Rai, 2019). We applied these rules to find probabilities of correlations between data points being highlighted. Show below:

	antecedents	consequents	antecedent support	...	lift	leverage	conviction
0	(BREAD)	(BISCUIT)	0.65	...	0.879121	-0.0275	0.938889
1	(BISCUIT)	(BREAD)	0.35	...	0.879121	-0.0275	0.816667
2	(COCK)	(BISCUIT)	0.15	...	1.904762	0.0475	1.950000
3	(BISCUIT)	(COCK)	0.35	...	1.904762	0.0475	1.190000
4	(COFFEE)	(BISCUIT)	0.40	...	0.714286	-0.0400	0.866667

Based on the combination relations, probabilities of support and confidences are then calculated. ‘Support’ helps indicate the frequency of the ‘if’ ‘then’ correlation within the combinations present in the data set. While ‘Confidence’ shows how often these respective correlations have

been shown to be true within the data set. These correlations then graphed to indicate the relationship between the confidence and support of the data points in this association analysis.

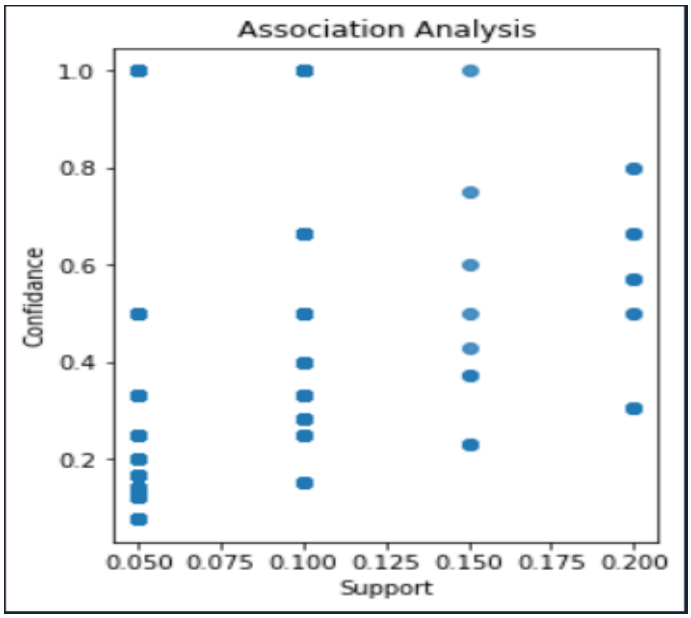
For clustering analysis we did hierarchical clustering and used ‘Agglomerative Clustering’. At first we visualized the dataset and as we do not require labels in clustering we dropped the ‘class’ label as the other four were numerical. To help us find any relation between the attributes we used pairplot, and the results are shown below:



The resulting pairplot clearly shows that there is a positive correlation between petal length(p_length) and petal width(p_width). Then we used only these two attributes to plot the data.

Result:

For our Association analysis, having calculated the “Confidence” values for each set of antecedent and consequent points, we plotted them against the “Support”. This helps see how frequently an item was bought together, how often this relationship is true as well as shows the strength of the relationship. The Association Analysis, Support vs Confidence figure is plotted below:

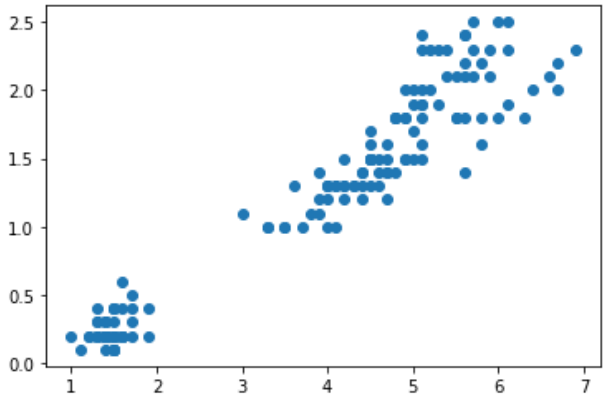


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(MILK)	(BREAD)	0.25	0.65	0.2	0.800000	1.230769	0.0375	1.75
1	(SUGER)	(BREAD)	0.30	0.65	0.2	0.666667	1.025641	0.0050	1.05
2	(CORNFLAKES)	(COFFEE)	0.30	0.40	0.2	0.666667	1.666667	0.0800	1.80
3	(SUGER)	(COFFEE)	0.30	0.40	0.2	0.666667	1.666667	0.0800	1.80
4	(MAGGI)	(TEA)	0.25	0.35	0.2	0.800000	2.285714	0.1125	3.25

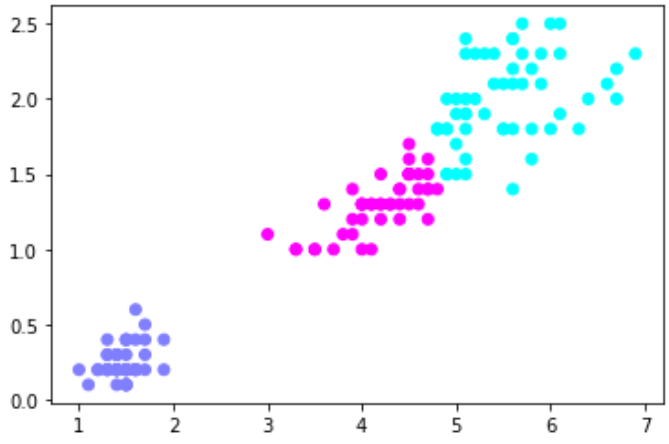
This tells us that when looking at the first row at index value 0, probability of milk sale occurring is 25% while the consequent support being bread is 65%. Based on the confidence value, we can indicate that 80% of customers that purchase milk will also purchase bread. While the leverage

indicates the difference in probability of consumption between the set antecedents and consequents.

For clustering analysis using the attributes petal length(p_length) and petal width(p_width) we made a scatter plot shown below:



Finally we use Agglomerative Clustering on the data for our clustering analysis. As there were three different classes in the original dataset we used 3 clusters and used euclidean for affinity. To better visualize the three clusters we did another scatter plot but this time using three different colors for the three different classes.



Conclusion:

From the results of the association analysis conducted on the list of grocery items purchased, we are able to see relationships of purchasable items. Using confidence and support values of the antecedents and consequences, we can conclude the likelihood of the set of items being purchased together. When looking at the clustering analysis, as the scatter plot shows two clusters are near each other. Those are the classes Iris Versicolour and Iris Virginica. While one of the clusters is at bottom left, which is the class Iris Setosa.

References:

1. Udwadia, S. (2016, November 8). *Grocery Store Data set*. Kaggle. Retrieved April 12, 2022, from https://www.kaggle.com/datasets/shazadudwadia/supermarket
2. Dhaduk, H. (2021, July 16). *Most powerful python functions apply() and lambda()*. Analytics Vidhya. Retrieved April 10, 2022, from https://www.analyticsvidhya.com/blog/2021/07/most-powerful-python-functions-apply-and-lambda/
3. Fisher, R. A. (1988). UCI Machine Learning Repository: Iris data set. Retrieved April 12, 2022, from https://archive.ics.uci.edu/ml/datasets/Iris
4. Rai, A. (2019, June 4). *Association rule mining: An overview and its applications*. upGrad blog. Retrieved April 10, 2022, from https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/