Bank Marketing Clustering Classification

CPS803 – Machine Learning

Aayush Regmi

500924277

1. Introduction

For this report I have chosen the Bank Marketing dataset which goes over the data accumulated by a Portuguese banking institution to determine whether a client is subscribed to their product (bank term deposit) or not. This data can be obtained from the UCI Machine Learning Repository. In this dataset, there are 16 relevant attributes (age, job, education, etc.) that will be used in classifying each client under the group of “subscribed” or “not subscribed”. The goal is to choose an appropriate clustering method that is able to accurately predict the result of whether an unknown client is subscribed (or going to subscribe) to the term deposit based off their given attributes.

I decided on using this dataset because it has plenty of instances, which would ensure that my clustering would be more robust as it would have plenty of data points to reference. Additionally, the dataset has a lot of categorical and binary attributes, as well as unknown values – which would allow me ample opportunity to comb through and pre-process the data in the best way that I see fit. Furthermore, the data as a whole is pretty interesting as it is a glimpse into the world of marketing and how machine learning can be a valuable tool.

2. Methods

In order to evaluate the data using a clustering method, I decided to choose K-means clustering as it felt like an ideal clustering method given the dataset. The basis of K-means clustering is choosing K different initial starting center points (centroid), assigning each point a category based off its closest centroid, and then adjusting the centroid based on the new center that is formed after assigning each point a category and repeating the process until the centroid no longer changes. Diagram 1 provides a visual representation of the process of K-means clustering on a dataset where we want 3 distinct classes. Likewise, in my dataset I will be using K-means clustering with 2 distinct classes (K=2), to help predict the clients that have subscribed vs have not subscribed.

In addition, I will also need to pre-process the data so that it can be interpreted and analyzed to create clusters. This includes changing all categorical attributes to numerical attributes using different encoding methods and removing any instances with unknown values. To change the categorical attributes into numerical attributes, I used binary encoding to represent the attributes (default, housing, and loan) that had only two values (yes or no). In this case, I encoded the value “yes” to equal 1, and “no” to equal 0. Otherwise, for the remaining categorical attributes I used the one-hot method (Diagram 2) to represent their values in a numerical manner.

To evaluate the result of the clusters, I will compared the classifications that are attributed to each of the data points based off the clustering with the actual classification of the data points. The criteria that I will use to evaluate these clusterings will be accuracy, precision, recall, f1-score, and the sum of

square differences between the data points and the centroid. These values should be able to give a good estimate of how well the machine has clustered the data and how representative it is of the actual results. The formula for calculating the the evaluation methods can be found below.

Recall = TP / (TP+FP)

Precision = TP / (TP+FN)

F1 = (2*Precision*Recall) / (Precision + Recall)

Accuracy = (TP+TN) / (TP+TN+FP+FN)

TP = # of True Positions (Correctly classified as true)

TN = # of True Negatives (Correctly classified as false)

FP = # of False Positives (Incorrectly classified as true)

FN = # of False Negatives (Incorrectly classified as false)

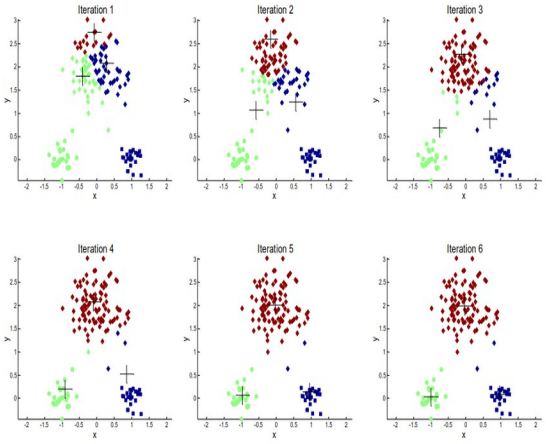

Figure 1: Example of K-means clustering

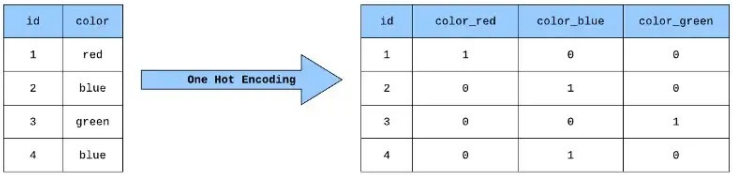

Figure 2: Example of one-hot encoding

3. Results

	Precision	Recall	F1-Score
Not subscribed	0.89	0.96	0.92
Subscribed	0.16	0.06	0.08

Accuracy = 0.86

SSE of samples to closest centroid = 192320776163.53

4. Conclusion

In conclusion, we can see that the overall accuracy of the clustering method is above 85%, which would mean that the clustering is fairly accurate with classifying the instances when compared to their actual classifications. However, the precision, recall and f1-score are very low for the subscribed classification, meaning that there were a lot of false positives and false negatives for subscribed clients that got incorrectly classified when creating the clusters. This could be a result of the imbalance between the two classes, as there were significantly more clients who were not subscribed compared to subscribed, so as a result the cluster may have been skewed more towards classifying the not subscribed clients compared to the subscribed clients. This can be seen by the high precision, recall and f1-score that the not subscribed instances were able to attain. Also, there will always be some level of variability each time a new clustering is generated by virtue of how the K-means clustering method

works. Overall, given the accuracy of the clustering, I believe that the model is still decent at correctly classifying the results. However, there is still probably some room for improvement.

5. References

Figure 1: Elodie Lugez D2L Slides (Chapter 6)
Figure 2: https://miro.medium.com/max/828/1*ggtP4a5YaRx6l09KQaYOnw.webp