K-Means Clustering

Anand Pandey   Katie Hidden   Akash Chandra

2022-11-05

1 Introduction

We have often heard of training machine learning models with labeled data. Imaging if there is a massive volume of data with no labels and we want to come up with a scalable approach to process these data and find insights. Difficult as it may seem, this is possible with clustering algorithms like K-Means.

Unsupervised machine learning is a type of algorithm that works on detecting patterns from a dataset when outcomes are not known or labeled. In unsupervised learning models it is not possible to train the algorithm the way we would normally do in case of supervised learning. This is because the data is neither classified nor labeled and allows the algorithm to act on the information without supervision. An unsupervised algorithm works on discovering the underlying hidden structure, pattern or association of the data and that helps the model in clustering or grouping the data without any human intervention.

The main goal of this paper is to discuss the concept and underlying methodology of one the unsupervised algorithm called K-Means clustering. We will also discuss how K-Means can be leveraged in real time applications like customer segmentation (Yulin 2020). In this paper, we will also be discussing various limitations and bottlenecks of K-Means clustering algorithm and would suggest some improved algorithms to overcome these limitations.

1.1 K-Means Clustering

K-means clustering is used for grouping similar observations together by minimizing the Euclidean distance between them. It uses "centroids". Initially, it randomly chooses K different points in the data and assigns every data point to their nearest centroid. Once all of them are assigned, it moves the centroid to the average of points assigned to it. When the assigned centroid stops changing, we get the converged data points in separate clusters.

There are some limitations of K-means clustering. It can be difficult to determine an appropriate initial K-value, especially with large and multidimensional datasets. The algorithm is sensitive to the initial centroid values and may fall into the local optimum solution. K-mean clustering typically classifies spherically shaped data well, but is less successful at classifying irregularly shaped data.

1.2 Customer Segmentation

Customer segmentation helps divide customers into different groups based on their common set of characteristics (like age, gender, spending habit, credit score, etc.) that helps in targeting those customers for marketing purposes. The primary focus of customer segmentation is to come up with strategies that helps in identifying customers in each category in order to maximize the profit by optimizing the services and products. Therefore, customer segmentation helps businesses in promoting the right product to the right customer to increase profits(Tabianan 2022).

Customer segmentation is not only helpful for business but also helps customers by providing them information relevant to their needs. If customers receive too much information which is not related to their regular purchase or their interest on the products, it can cause confusion on deciding their needs. This might lead their customers to give up on purchasing the items they required and effect the business to lose their

Table 1: Data Definition

Invoice	Nominal	Invoice number. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter ‘c‘, it indicates a cancellation.
StockCode	Nominal	Product (item) code. A 5-digit integral number uniquely assigned to each distinct product.
Description	Nominal	Product (item) name.
Quantity	Numeric	The quantities of each product (item) per transaction.
InvoiceDate	Numeric	Invice date and time. The day and time when a transaction was generated.
Price	Numeric	Product price per unit in sterling (£).
CustomerID	Numeric	Customer number. A 5-digit integral number uniquely assigned to each customer.
Country	Nominal	Country name. The name of the country where a customer resides.

potential customers. The clustering analysis will help to categorize the customer according to their spending habit, purchase habit or specific product or brand the customers interested in. Customer segmentation can be broadly divided into four factors - demographic psychographic, behavioral, and geographic(Tabianan 2022). In this paper, customer behavioral factor has been primarily focused.

K-Means clustering algorithm can help effectively extract groups of customers with similar characteristics and purchasing behavior which in turn helps businesses to specify their differentiated marketing campaign and become more customer-centric.(Yulin 2020)

2 Methods

3 Analysis and Results

3.1 Dataset Description

For the customer segmentation, we will be using a data set that contains all the transactions that has occurred for a UK-based non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers. The company was established in 80s as a storefront and relied on direct mailing catalogues and taken order over phone. In recent years, the company launched a website and shifted completely to a web based online retail to take technological advantage of customer-centric targeted marketing approach(Chen 2010).

The customer transactions dataset has 8 variables as shown in below data definition table.

3.2 Data Preparation

Before we prepare the data, let’s take a look at the summary and see if there are any missing values or other trends.

The CustomerID contains unique Id for each unique Customer. However, we have 107,927 customers with no customer ids. We will be removing these customer before starting our analysis.

Table 2: Summary of Data

Quantity	525461	0	10.34	107.42	-9600.00	1.00	3.0	10.00	19152.00
Price	525461	0	4.69	146.13	- 53594.36	1.25	2.1	4.21	25111.09
Customer Id	417534	107927	-	-	-	-	-	-	-

```
df=na.omit(df, cols="CustomerID")
```

3.3 Data and Vizualisation

3.4 Statistical Modeling

4 Conlusion

References

Chen, Dr. Daqing. 2010. "Online Retail II Data Set." School of Engineering, London South Bank University, London SE1 0AA, UK. https://archive.ics.uci.edu/ml/datasets/Online+Retail+II.

Tabianan, S., Velu. 2022. "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data." *Analytical Methods* 14 (12): 7243. https://doi.org/10.3390/su14127243.

Yulin, & Qianying, D. 2020. "A Study on e-Commerce Customer Segmentation Management Based on Improved k-Means Algorithm." *Information Systems and e-Business Management* 18 (4): 497–510. https://doi.org/10.1007/s10257-018-0381-3.