Indian Institute of Technology, Indore

Department of Computer Science and Engineering

CS 304 Project Proposal
Customer Segmentation: An Approach based on K-Means Algorithm

1. Purnadip Chakrabarti (190002048)
2. Vaibhav Chandra (190001065)

Submitted to: Dr. Aruna Tiwari

January 16, 2022

Introduction

Customer Segmentation is a very crucial problem for modern businesses in the process of making various marketing policies. Different groups of customers react differently to various kinds of policies proposed by the businesses. Hence, the businesses have to come up with different policies for different classes of it's customers so that more revenue is generated.

In earlier days, this process of targeting specific groups of people was done manually. But today, when there are millions of customers of a company all around the globe, it is practically impossible to do the same manually. So, to do the same task for such a large amount of data, machine learning algorithms are used to allow businesses to make an informed decision about their future course of action.

Problem Statement

We are given a data set which contains data regarding many transactions that have taken place over a course of time. The data is in the form of comma separated values in which each row consists of various attributes pertaining to one transaction. Product involved, country in which the transaction took place, Customer ID are some of these attributes.

We are supposed to analyze the purchasing habits of the customers and try to divide them into different groups on the basis of many factors like the type of product purchased, price of the product purchased, etc. After the segmentation of the customers we are also supposed to calculate the inertia and Dunn index of the clusters to report the performance of the algorithm. A point to be noted is that this is a problem related to Unsupervised Learning.

Data Description

The dataset that we have chosen can be found in this link: E-commerce dataset
This data contains details of 540,909 transactions made on an E-commerce website over a course of 1 year. Each transaction includes the following details:

1. Customer ID
2. Invoice Number
3. Stock Code
4. Item Description
5. Quantity of item
6. Invoice Date
7. Price per unit of item
8. Country

We are planning to use Item Description, Quantity, Price per unit of item, country in our model. We will be incorporating other columns too if some customer related information is found there. We won't use Customer ID because apparently that doesn't play any role in classifying customers. Moreover, all the non-numeric data will be transformed into a numeric form and then fed into the model. However, these decisions are subject to change depending on the characteristics of the data-set.

Our Approach

For the clustering of the transaction data, we will be employing K-Means algorithm, a rather well known algorithm in the area of unsupervised learning, used to recognize clusters in unlabelled data. We went forward with this particular algorithm for now but we may choose to use other clustering algorithms in future if they prove to be a better prospect.

K-means takes K as an input parameter where K is the number of clusters that we want to have after segmentation. We begin by taking K random points as the centroids of K clusters initially. We then assign each data point to exactly one of these K clusters. After the formation of these clusters, we calculate the centroids of these clusters. These centroids we obtain are the new K centers of the clusters. We repeat this process repeatedly until there is no more segmentation or a pre-decided number of iterations are reached.

Our Inspiration

While everyone, who were enrolled in this course, were taking up supervised learning for their minor project, we chose to investigate a relatively unexplored area of unsupervised learning.

We found that clustering of data is a prominent area in this domain and some of the well known applications of clustering are customer segmentation, grouping of similar songs and many more. Since both of us are inclined towards learning about the businesses, customer segmentation was aligned with our interest and hence we chose to go forward with this project and solve this real life problem.