GRADUATION PROJECT

CUSTOMER SEGMENTATION AND ENRICHMENT USING EXPECTATION MAXIMIZATION ALGORITHM

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
Chair of Stochastic Operations Research

Niveditha Kumar (s2030373)

Graduation Committee:
Prof. dr. Richard Boucherie (University of Twente)
Dr. ir. Jasper Goseling (University of Twente)
Dr. Julio Backhoff (University of Twente)
Dr. Patrick de Oude (Albert Heijn)

February 2020

UNIVERSITY OF TWENTE.

Preface

This report is the result of my graduation project at University of Twente. The project was carried out in collaboration with the Data Science department at Albert Heijn (AH). I would like to take the opportunity to thank the people who are responsible for playing an instrumental role in this project.

Firstly, I would like to thank Jasper Goseling without whom this thesis would not have seen its end. I am grateful for his support when I hit roadblocks, feedback on my progress and report and guidance in helping me understand the mathematical content better. I have learnt a lot throughout the period of the graduation project. Secondly, I would like to thank Patrick de Oude for giving me the opportunity to work at AH. I am thankful for the inputs, feedback and the possibilities given in order to contribute and improve my data science skills. Furthermore, I would also like to thank my colleagues at AH for helping me with understanding and navigating through the data sets. I would also like to thank Mark for his support, encouragement, help and for being there for me when I needed it. I would also like to thank my family and friends for always being one call away.

Abstract

Targeting customers based on their interest in order to show personalised advertisements helps companies to improve their revenue. In the retail setting, for showing personalised advertisements customer segmentation is done. Existing customer segmentation is carried out at Albert Heijn (AH) based on business rules.

Existing customer segmentation at AH makes use of only a few features. Manually adjusting the threshold based on a feature can add customers that are not necessarily interested in a product into the segment. The goal of the project is to add additional customers to existing customer segments based on the Expectation Maximization (EM) algorithm. Enrichment of the segment is done by adding additional features. It is assumed that customers that are interested in organic products and other customers are drawn from two different Gaussian distributions. The customers are mathematically represented as realizations of random variables in a sample space consisting of all the features and underlying class labels. Based on the mathematical formulation and assumptions, the algorithm helps to classify if a customer is interested in organic products or not.

In order to show that adding additional features does help to enrich the segment, numerical experiments are carried out on the synthetic data set. The EM algorithm suffers from locally optimal solutions and the final parameter estimates depend on the starting parameters. So, numerical experiments are set up with different initialization of the parameters of the distribution. Based on the numerical experiments a number of observations is highlighted. The first observation is that adding additional features does help to enrich the segment. It is also observed that the EM algorithm fails to differentiate between completely overlapping clusters. The last observation is that the EM algorithm works best when initialized based on business rules.

The EM algorithm is further applied on the AH data set. Based on visualising the data it can be observed that that data points are not linearly separable. Similar to the numerical experiments, it is observed that when features that are highly correlated to the organic segment are added, the algorithm is able to decrease type 1 and type 2 error and capture most of the relevant customers in order to enrich the segment. When features that are less correlated to the organic segment are added, the algorithm gives a higher type 1 and type 2 error.

The problem was formulated mathematically and a framework is provided in order to segment and enrich existing customer segments. It can be concluded that adding additional features that are correlated with the organic customers based on business rules helps to enrich the segment. The EM algorithm is set up which is a soft clustering algorithm and the output is easy to interpret. Further improvements with respect to initialization of the parameters, can be done based on the results obtained via A/B testing. A clustering based recommender system is also recommended as future research.

Keywords: customer segmentation, Expectation Maximization algorithm, A/B testing

Contents

Chapter 1

Introduction

This graduation project was carried out at Albert Heijn (AH). AH is the largest Dutch supermarket chain, founded in 1887, it has a market share of over 30%. There are various departments such as Finance, Supply Chain and Data Science at AH. This graduation project was carried out at the Data Science department that deals with strategy and analytics. The team at AH have developed rules to group customers (customer segments) based on the products they buy. Based on the segment a customer belongs to, advertisements (ads) are shown to customers on the AH website.

Customer segmentation is the division of customers into similar groups based on shared characteristics, purchasing behavior or consumption patterns. Segmenting customers helps in targeting customers based on their interest and advertise to them effectively [1]. In order to target customers and show them personalised ads, most companies incorporate the following ways for segmentation:

- No segmentation, show ads to all customers;
- Rule based segmentation, performed by human analysts;
- Automated segmentation, using machine learning algorithms [2].

Existing customer segmentation at AH is rule based, referred to as business rules. Customer segmentation is used to show ads, see Figure 1.1, to customers interested in the products based on customer segmentation. By showing relevant ads, the likelihood of clicking on the ad and buying the product increases, showing ads to customers who are not interested in the product is a lost opportunity to show them ads based on their interest.

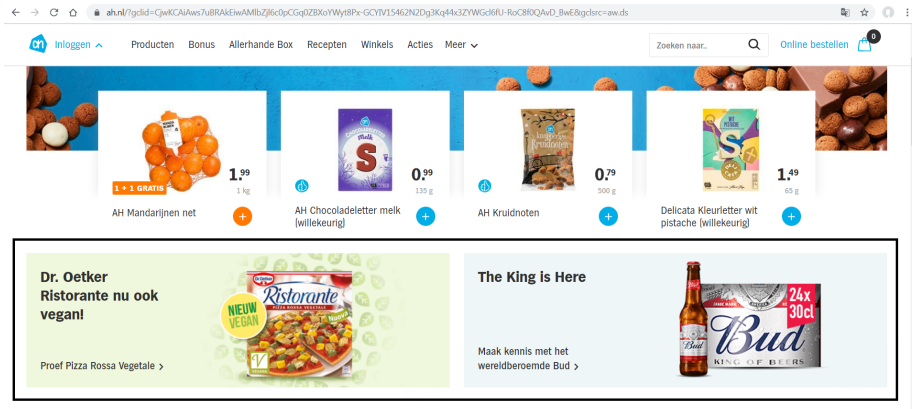

Figure 1.1: Albert Heijn ads for customers interested in Vegan Pizza and Beer

Depending on the number of clicks on the ad, the supplier of the product pays AH. Improving targeted marketing, showing more personalised and relevant products to each customer based on their interest, increases the probability of profit due to targeted marketing. Therefore, the goal is to add

additional customers who are interested in the product to the existing customer segments, known as customer enrichment. Adding customers that are not interested in the product is a lost opportunity to show them ads based on their interest.

1.1 Problem Description

AH uses rule based segmentation (business rules) in order to segment customers. The existing customer segments found based on the business rules at AH are for example, Lipton ice tea, organic products, baby products and customers sensitive to promotions. In order to describe the business rules, an use case is selected and used throughout the report. The use case selected is customers who are interested in organic products. First, customers who are interested in organic products are found based on filtering for customers who have bought organic products in the last 90 days. This filtering of customers is based on both online and offline transaction data. Online transaction data refers to the customers that buy products online at the AH website. Offline transaction data refers to the data that customers buy at the AH stores.

The business rules used to segment organic customers is defined mathematically in Equation (1.1). The business rules is applied after filtering for the period of 90 days. The last 90 days are considered due to the fact that online data of customers can only be stored for a period of 90 days, as per data protection policies. The business rule used to segment organic customers is defined below,

$$T_i > \mu_T \ , P_i > \mu_0, \tag{1.1}$$

where for each customer i, $i = 1...N$;

- T_i refers to the count of total organic products bought (total products) by each customer i;
- μ_T is the mean of total products;
- P_i refers to the count of unique organic products bought (organic products) by each customer i;
- μ_0 is the mean of organic products.

Equation (1.1) states customers who buy total products greater than mean of total products bought or who buy unique organic products greater than mean of unique organic products bought are considered organic customers. For example, when mean of customers who bought organic products equals 12 ($\mu_0 = 12$), customers who bought 12 or more organic products ($\mu_0 \geq 12$) in the last 90 days are labelled as organic. This business rule includes two features, count of total products (total products) and count of unique organic products bought (organic products). The business rules can be generalised for other use cases as well.

The goal of this project is to add additional customers to the existing customer segments, also known as enrichment. In Figure 1.2, based on business rules, plot of the count of the unique organic products bought is seen. In Figure 1.3, plot of first two features can be seen. As per the Equation (1.1), in Figure 1.2, the line in black is the business rules which segments the customers as organic or other. Based on business rules, customers who have bought organic products above 12 ($\mu_0 >= 12$) are labelled as organic. The purple dots (True Positives, TP), represent customers who are interested in organic products based on business rules, the yellow dots (True Negatives, TN) represent customers who are not interested in organic products. The blue dots (False Negatives, FN) are the customers who are interested in organic products, but have not been captured when the business rules are applied. The green dots (False Positives, FN) are the customers labelled as organic based on business rules but are not interested in organic products. Decreasing the threshold (μ_0) in 1.2 can help to add more customers into the organic segment but the customers added may not be interested in buying organic products. Therefore, the goal is to add additional customers to the existing segments such that the customers that are not interested in buying organic products are not included in the organic segment. Ads shown to customers that are interested in organic products helps to increase the likelihood of clicking on the ad and buying the product. Showing ads to customers who are not interested in the product is a lost opportunity to show them ads based on their interest.

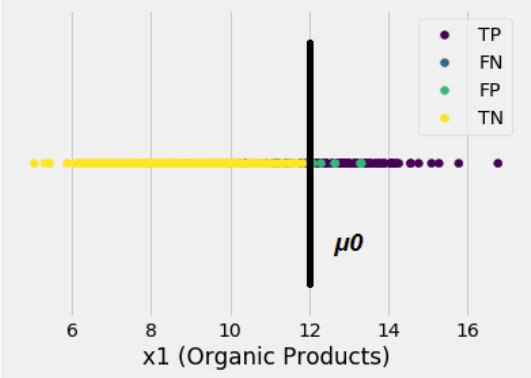

Figure 1.2: Plot of first feature

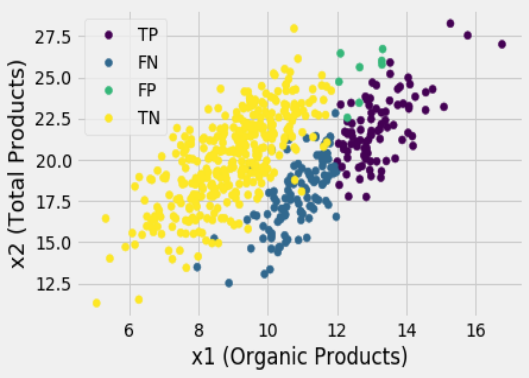

Figure 1.3: Plot of first two features

1.2 Thesis Contribution

First, in order to make the problem concrete a mathematical model is developed. The customers are defined as observations in a sample space. The observations are realizations of independent and identically distributed random variables with underlying ground truth and features. The problem is mathematically formulated and described in Chapter 3.

Currently the segmentation at AH is carried out manually. A machine learning approach in order to segment customers is set up. Based on the mathematical formulation, a Gaussian mixture model using the Expectation Maximization (EM) algorithm is set up in order to segment and enrich customers. The EM algorithm outputs a soft clustering algorithm. So, it reflects the probability with which each customer belongs to the organic and other segment. In order to initialize the parameters of the EM algorithm, initializing based on business rules is proposed. Existing method at AH based on business rules with the use of one feature and the proposed EM algorithm based on adding additional features in order to enrich customers are compared. The EM algorithm results in a more meaningful enrichment of customers, based on analysis of type 1 and type 2 error and proved to be a better approach than business rules inorder to enrich the segment.

In order to enrich existing customer segments the idea of adding additional features to find additional customers who are interested in organic products was put forth. The business rule for instance uses features $X_1....X_M$, additional features $X_{(M+1)}....X_D$ are added. The customers are mapped to a higher dimensional space. Based on numerical experiments on synthetic data set it can be seen that adding features helps to decrease type 1 and type 2 error and provide meaningful enrichment. Similarly, this approach is applied to the AH data set. However, clustering in a high dimensional feature space provides meaningless clusters, as most of the data points are clustered as part of the organic segment. This implies that, adding all the available features in high dimension provides a high type 1 error. In order to reduce type 1 and type 2 error, the best combination of features are selected based on correlation. Finally, the experiments on synthetic and AH data sets result in a framework. The developed framework provides enriched customer segments.

1.3 Thesis Overview

The organisation of the report is as follows. In Chapter 2 the related literature is discussed. In Chapter 3 the problem is formulated mathematically and the research questions are discussed. Also, the challenges faced are discussed. In Chapter 4 the approach, assumptions and the Expectation Maximization (EM) algorithm are discussed in detail. In Chapter 5 numerical experiments are carried out on synthetic data set and conclusions are drawn. In Chapter 6, the approach is applied to the AH data set. Finally, in Chapter 7, answers to the research question, conclusions and recommendations for future work are given.

Chapter 2

Related Work

In the literature review existing segmentation techniques based on machine learning approaches will be discussed with respect to algorithms, advantages, disadvantages and scalability of the model.

In customer segmentation, one of the approaches in order to recommend customers relevant products is incorporating recommender systems. In the next section existing approaches on recommender systems are mentioned.

2.1 Recommender Systems

Recommender systems recommend items to users based on the customers needs and preferences [3]. Recommender systems are broadly classified into content based, collaborative filtering and hybrid recommender systems. In content based recommender system the user is recommended items similar to the items preferred in the past. Collaborative filtering recommends items to users which have been liked by people with similar preferences. Hybrid recommender systems combine the first two approaches, collaborative and content based recommender systems. Recommender systems have the problem of the user being limited to getting recommendations for items that have only already been rated [4]. A lot of approaches propose a clustering based recommender system. First, clustering techniques are applied to identifying groups of users who appear to have similar preferences. Once the clusters are created, predictions for an individual can be made by averaging the opinions of the other users in that cluster [5].

When customer labels are known, supervised classification algorithms such as neural networks, latent discriminant analysis and decision-tree induction can be carried out [6]. The aim is to create a model such that the class labels are assigned to one of the classes. When there is information about class-labels for some data points, semi-supervised clustering can be done. Semi-supervised partitional algorithms and hierarchical clustering are discussed in [7]. With the AH data set there are no clear class labels assigned. When there are no clear labels assigned to the data, unsupervised clustering methods are used to label the data. The purpose of clustering is to identify patterns and similarity with data points and clusters [8]. In the next section, discussion is carried out on unsupervised clustering approaches.

2.2 Unsupervised Clustering

Clustering can broadly be divided into partitional clustering, density based clustering, hierarchical clustering and model based clustering. Different clustering algorithms and its applications are discussed below.

2.2.1 Partitional Clustering

Partition clustering algorithms partition data points into K clusters and each partition represents a cluster. In partitional clustering, each point belongs to only one cluster and each cluster consists of at least one point. Fuzzy partition is an exception where each data point can belong to more than one cluster [9]. Most partitional algorithms such as K-means and K-mediods minimize the following objective function:

$$D(x_i, \mu_k) = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} |x^i - \mu_k|^2, \tag{2.1}$$

where w_{ik} = 1 if the point belongs to the cluster and 0 otherwise and μ_k is the centroid of the kth cluster. The general algorithm of K-means is:
1. Select K points as initial centroids
2. Repeat step 1
3. Form k clusters by assigning all points to the closest centroid
4. Recompute the centroid of each cluster until the centroids do not change

There are different papers that discuss application of K-means for customer segmentation [10] [11] [12]. K-means is one of the most commonly used algorithms. The advantages of K-means is that it is easy to interpret and is scalable. One of the major drawback of the algorithm is that it cannot handle non-globular structures. Also, Kmeans do not have the capability to separate clusters that are non-linearly separable in input space [13].

There are modified versions of K-means such as FORGY, ISODATA, CLUSTER and WISH [14]. There are different ways to compute the optimal number of clusters (K) such as Elbow Method, Average Silhouette Method [15] and Gap Statistic Method [16]. There are different distance measures that are used in clustering and the commonly used distance metrics are discussed in Section 2.4. The most commonly used distance measure is the Euclidean distance.

Fuzzy C-means (FCM) is also a centroid based algorithm. The algorithm assigns membership value to each data point between 0 and 1 to indicate the belongingness to each of the K clusters. This is a type of soft clustering whereas K-means is a hard clustering algorithm. The other clustering algorithms based on FCM are PAM, CLARA and CLARANS [17]. Like K-means, fuzzy C-means also models the clusters as spheres. The FCM ensures good convergence guarantees. The FCM takes a long time to converge for large data sets [18].

Affinity propogation (AP) is a relatively new algorithm proposed in 2007 [19]. Unlike most partitional clustering algorithms which refine and store a fixed number of cluster centers, AP algorithm regards all data points as potential cluster centers. The disadvantage of this algorithm is that it is not suitable for very large data sets, and the clustering results is sensitive to the parameters involved in the AP algorithm [20].

2.2.2 Hierarchical Clustering

Hierarchical clustering take a hierarchical approach to find clusters in the data. Broadly, they are divided into two types: agglomorative and divisive clustering. Agglomorative clustering is a bottom up method of clustering where each data point is considered as a cluster and similar data points are merged recursively to output one cluster. Divisive clustering is when all the points are considered as one cluster and are recursively divided into smaller clusters.

The basic idea of hierarchical clustering is that you have 'n' groups and end up with 1 group or vice versa. First, distance between each point is calculated. The points that have very low distance between each other are bought together as a group. Then the distance between the created group and data points are calculated. The common distance metrics used for this purpose are average linkage, complete linkage and single linkage. Average linkage chooses average distance from cluster to a point outside the cluster. Complete linkage chooses the longest distance from any member of one cluster to any member of the other cluster. In single linkage, the minimum point is chosen. Unlike

partition clustering algorithms, the number of clusters need not be selected apriori. But, the major drawback of hierarchical clustering is that once the two points are linked, they cannot be linked to any other hierarchy. So, if two data points are linked and the linkage is not good, it cannot be undone. Another disadvantage of agglomorative clustering is that it is not scalable [21].

Some of the clustering algorithms based on agglomorative clustering approaches are BIRCH [22], CURE [23], ROCK and CHAMELEON [24]. BIRCH and CURE handle outliers well. BIRCH has a better time complexity but lacks in cluster quality than CURE algorithm. ROCK uses an agglomorative clustering approach for categorical data set. Common applications of hierarchical clustering algorithms are in the field of social sciences and biological taxonomy.

2.2.3 Density Based Clustering

Density based clustering is used to identify clusters of arbitrary shape. This is based on the fact that the density of data points in a cluster is higher than outside the cluster. The less dense regions are recognized as noise. The most commonly known algorithm is DBSCAN (Density Based Spatial Clustering of Applications with Noise) [25]. User needs to specify the radius and minimum number of objects it should have in its neighborhood as input parameters. Although DBSCAN has the ability to identify clusters of arbitrary shape, it is sensitive to outliers and parameters. DBSCAN does not perform well if the data is high dimensional and euclidean distance is used to find proximity of objects. DBCLASD and OPTICS are an extension to DBSCAN [26]. DENCLUE is an aggregate of partitioning and hierarchical clustering approaches [27].

2.3 Model Based Clustering

Model based clustering attempts to fit the given data into a mathematical model. The resulting cluster in model based clustering offers better interpretability than similarity based approaches [28]. It assumes that data are generated by a underlying probability distribution. Each cluster corresponds to a different distribution, and in general the distributions are assumed to be Gaussian [29].

The Expectation Maximization (EM) algorithm is an iterative algorithm where a single iteration is composed of the E and the M step which finds the maximum likelihood of the model [30]. It has been applied to many statistical learning problems such as Gaussian Mixture Models (GMM) and hidden Markov models [31]. The EM algorithm is widely used as it is stable and is mathematically proven to converge [32]. Let $X = \{x_1, x_2 ... x_N\}$ be the data points, let Z be the underlying ground truth with class labels 0 or 1. It is given as $Z = \{z_1 ... z_N\}$ and is commonly referred to as latent variables or unobserved variables. Let θ be the model parameters. $\{X, Z\}$ is referred to as the complete data set and X as the observed data set. The goal is to maximize the observed log likelihood,

$$\ell(\theta) = \sum_{n=1}^{N} \ln \left\{ \sum_{z_n} \mathbf{P}\left(x_n, z_n | \theta\right) \right\}. \tag{2.2}$$

Maximizing the observed log likelihood is not an easy task due to the presence of summation inside the log. The log likelihood of the complete data set is given by,

$$\ell_c(\theta) = \sum_{n} \ln \mathbf{P}\left(x_n, z_n | \theta\right). \tag{2.3}$$

Since the complete data set is not observed, it is also not possible to maximize the complete log likelihood $\ell_c(\theta)$. The basic idea of the EM algorithm is to find the maximum $\ell(\theta)$ via a two step process. The expected complete log likelihood is calculated with respect to the conditional distribution of Z, given observed data and the current estimate of the parameter θ. The conditional expectation is denoted as $\mathcal{Q}\left(\theta, \theta^{\text{old}}\right)$. This is the expectation step.

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \mathrm{E}_{\mathrm{Z|X},\theta^{\text{old}}}(\ell_c(\theta)) = \sum_{\mathbf{n}} \sum_{\mathbf{z_n}} \mathbf{P}\left(z_n | x_n, \theta^{\text{old}}\right) \ln \mathbf{P}(x_n, z_n | \theta) \tag{2.4}$$

In the next step, the values of the parameters are updated by maximizing $\mathcal{Q}(\theta, \theta^{\text{old}})$.

1. Choose initial setting for parameters θ^{old}
2. **E step** Evaluate $\mathcal{Q}(\theta, \theta^{\text{old}}) = \mathrm{E}_{\mathrm{Z|X},\theta^{old}}(\ell_{\mathrm{c}}(\theta)) = \sum_n \sum_{z_n} \mathrm{P}(z_n|x_n, \theta^{\text{old}}) \ln \mathbf{P}(x_n, z_n|\theta)$
3. **M step** Evaluate θ^{new} given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\theta} \mathcal{Q}\left(\theta, \theta^{\text{old}}\right) \tag{2.5}$$

4. Check for convergence of either the log likelihood or parameters values. If convergence criteria not yet satisfied then

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}} \tag{2.6}$$

and return to step 2.

For multimodal distributions the EM algorithm can converge to a local maxima of the observed likelihood function depending on the starting values. The EM algorithm has the important property that the log likelihood of the observed data increases at each step. EM only converges to a stationary point, which may not be a local maxima and not much is known about the rate of convergence [32]. The properties of the EM algorithm is discussed in detail in [33].

Overlapping clusters is something that is not tackled often in clustering algorithms. In [34], the issue of tackling overlapping clusters via the EM algorithm is discussed. The most common applications of the EM algorithm is image segmentation [35]. Other common applications of the EM algorithm is in data clustering, parameter estimation for mixed models [36], especially in quantitative genetics [37]. The EM algorithm is also used in medical image reconstruction [38], robotics [39] and signal processing [40].

COBWEB is also a model based clustering algorithm that creates a hierarchical tree in the form of a classification tree [41]. CLASSIT is an extension of COBWEB that allows mixed nominal and numerical attributes [42]. SOM (Self - Organising Maps) maps all points in a high dimensional space in a 2-3 dimensional space such that distance and proximity are preserved as much as possible [43].

[44] provides a comprehensive survey of the clustering algorithms. In the next section, the commonly applied distance metrics that are key to clustering are discussed.

2.4 Distance Metrics

In clustering the most common used distance measures are given below and papers related to the distance measures in [45]. In higher dimension $(d > 10)$ the curses of dimensionality comes into the picture [46]. The distance measures can become meaningless as points can be equidistant from each other [47]. The choice of distance measure also depends on that features used in the data set. Given data points x_i and x_j, the various distance measures are defined as follows.

1. Euclidean distance
Euclidean distance performs well with compact and isolated clusters [48]. Kmeans is implicitly based on pairwise Euclidean distance between points because Equation (2.1) can be written as the sum of pairwise squared Euclidean distances divided by the number of points. One drawback of Euclidean distance is that the largest-scaled features dominates the other [49]. This can be overcome by normalizing the features. The Euclidean distance between two points, x_i and x_j in the l^{th} dimension is given as:

$$D\left(x_i, x_j\right) = \left(\sum_{l=1}^{D} \left|x_{il} - x_{jl}\right|^2\right)^{\frac{1}{2}}. \tag{2.7}$$

2. Manhattan or City block distance
Manhattan distance is sensitive to outliers. When this distance measure is used, the shape of the

cluster is hyper-rectangular [50]. One advantage of Manhattan distance over Eucledian distance is the reduced computational time. The distance is defined as:

$$D\left(x_i, x_j\right) = \sum_{l=1}^{D} \left|x_{il} - x_{jl}\right|. \tag{2.8}$$

3. Mahalnobis distance
The Mahalnobis distance takes into account the covariance of the features. It is based on the correlation between variables because of which different patterns can be identified and analysed [51]. The distance is defined as:

$$D\left(x_i, x_j\right) = \left(x_i - x_j\right) S^{-1} \left(x_i - x_j\right)^T, \tag{2.9}$$

where S^{-1} is the covariance matrix.

4. Cosine distance
Cosine distance measures the cosine of the angle between the vectors. It is most commonly used to measure document similarity in text analysis [52]. The cosine distance is defined as:

$$Cosine\left(x_i, x_j\right) = 1 - \left(\frac{x_i^T x_j}{\left\|x_i\right\| \left\|x_j\right\|}\right). \tag{2.10}$$

In the literature review, various ways to segment customers was discussed along with its advantages and disadvantages. A discussion on the most commonly used distance metrics was also carried out. The clustering algorithms are weighed with respect to the AH data set. In the next section, the problem is formulated mathematically, the research questions and ways to evaluate the model are put forth.

Chapter 3

Mathematical Formulation

Based on the problem description, the business rules uses features $\mathcal{F}$, where $\mathcal{F}$ = $\{X_1....X_M\}$. The goal is to enrich the existing customer segments. In order to enrich the existing customer segments at AH, adding additional features is proposed. Additional features $\mathcal{A}$, where $\mathcal{A}$ = $\{X_{(M+1)}....X_D\}$, are added in order to enrich the segment, where $D > M$. Let $\mathcal{F} \times \mathcal{A} = \mathcal{X}$ where $\mathcal{X} \in \mathbb{R}^D$, a D - dimensional vector space with $\{X_1, X_2....., X_D\}$. There is an underlying ground truth feature $\mathcal{Z}$, where customers are labelled as organic (0) and other (1). This underlying ground truth is not known with the AH data set. The mathematical formulation is developed based on this approach.

Consider the sample space $\mathcal{L}$ that consists of all the features $\mathcal{L}$ = $\mathcal{Z} \times \mathcal{F} \times \mathcal{A}$ where,

- $\mathcal{Z}$ is the feature that consists of class label $\{0, 1\}$, where 0 represents if a customer belongs to the organic segment and 1 if the customer belongs to the other segment;
- $\mathcal{F}$ are the continuous features used in the business rules;
- $\mathcal{A}$ are the additional continuous features that are added.

The ratio of the points labelled as other and organic customers is defined as;

$$w_0 = P(Z = 0), \tag{3.1}$$

$$w_1 = P(Z = 1). \tag{3.2}$$

The probability of the attributes belonging to class label 0 and 1 can be defined as;

$$P_0(f, a) = P(F \leq f, A \leq a \mid Z = 0), \tag{3.3}$$

$$P_1(f, a) = P(F \leq f, A \leq a \mid Z = 1). \tag{3.4}$$

N observations, $\{o_1....o_N\}$ are drawn from the sample space $\mathcal{L}$ and each $o_i = (z_i, f_i, a_i)$. It is assumed that the observations in the sample space are realizations of independent and identically distributed random variables. $\mathcal{T}$ is the underlying ground truth in the data set that can be defined as;

$$\mathcal{T} = \{(z, f, a) \in \mathcal{L} : Z = 1\}. \tag{3.5}$$

Let ψ be the function that is a mapping $\psi : \mathcal{F} \rightarrow \{0, 1\}$, that outputs labels 0 or 1 based on the features used in the business rules. Then, let $\mathcal{B}$ be the set that is defined based on business rules as;

$$\mathcal{B} = \{(z, f, a) \in \mathcal{L} : \psi(f) = 1\}. \tag{3.6}$$

Let ϕ be a function that is a mapping $\phi : \mathcal{X} \rightarrow \{0, 1\}$, that outputs 0 or 1 based on the algorithm developed, where the features from business rules and additional features are included. Let $\mathcal{S}$ be the enriched set, which also includes $\mathcal{B}$. $\mathcal{S}$ are the customers that we want to find,

$$\mathcal{S} = \{(z, f, a) \in \mathcal{L} : \phi(x) = 1\}. \tag{3.7}$$

Ideally, we would like $\mathcal{S} = \mathcal{T}$, the customers that we want to target. The mathematical formulation is shown in the Figure 3.1. Based on the mathematical formulation, a way to evaluate the type 1 and type 2 error of the algorithm is set up in the next section.

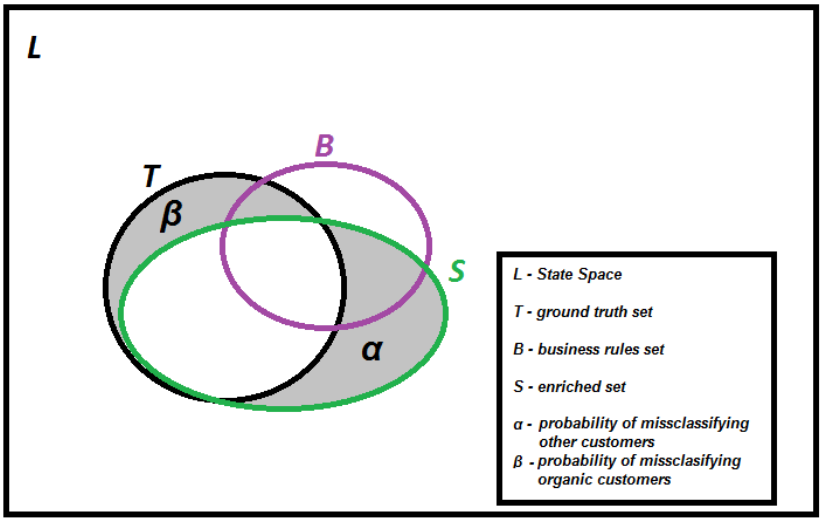

Figure 3.1: Representation of the sample space

3.1 Type 1 and Type 2 error

The algorithm $\phi(\cdot)$ is a mapping $\phi : X \to \{0, 1\}$ that returns a predicted binary label 0 or 1 given X. Classification error occurs when $\phi(X) \neq Z$. Then the probability of missclassifying can be defined as

$$\alpha = P(\phi(X) = 0 \mid Z = 1) = P_1(\phi(X) = 0), \tag{3.8}$$

$$\beta = P(\phi(X) = 1 \mid Z = 0) = P_0(\phi(X) = 1). \tag{3.9}$$

α is the conditional probability of misclassifying a class 0 observation as class 1 and β is the conditional probability of misclassifying a class 1 observation as class 0. The type 1 and type 2 errors are represented in a confusion matrix as seen in 3.2. Type 1 error is when customers from the other segment are labelled as organic (FP) and type 2 error is when customers from the organic segment are labelled as other (FN). The goal is to enrich the customer segment by increasing type 2 error and decreasing type 1 error with respect to the business rules.

		Actual	
		Organic (0)	Other (1)
Predicted	Organic (0)	True Postive (TP)	False Positive (FP) (Type 1 Error α)
	Other (1)	False Negative (FN) (Type 2 Error β)	True Negative (TN)

Figure 3.2: Evaluation

Based on the confusion matrix, recall and precision can be analysed. Recall and precision is

given in (3.10) and (3.11).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.11}$$

Precision talks about how precise the model is out of the predicted positive outcomes. A high precision relates to a low false positive rate. Recall calculates how many of the customers that are organic are actually captured. A high recall pertains to a low false negative rate.

3.2 Challenges

There are a number of challenges faced with this problem. The first problem is that the data points may not be linearly separable and the clusters can be arbitrarily shaped.

The second challenge is the curses of dimensionality [46]. The data points in high dimension (d$>$10) can be equidistant from each other. The discrimination of the nearest and farthest point in particular becomes meaningless. So, the choice of the distance metric is key. This can also have an impact on the type 1 and type 2 error.

The subspace for the organic customers is not well-defined. There can be cases where customers buy organic products by mistake or buy organic products for someone else. Such situations cannot be inferred from the available data.

Availability of the required AH data set is a challenge. Due to General Data Protection Regulation (GDPR) [53] on cookie consent, it is difficult to get access to all the customer data. The click data of customers is currently not available. Click data provides details on the number of clicks on products on the AH website when ads are shown. Based on the click data, customers can be labelled as organic or other customers. This can be used as the ground truth labels. There are also privacy concerns with getting details of customers from third-party companies. Using of the third-party data of customers with AH's own data set, for which the customer did not consent, can be a cause of concern.

Based on literature review, mathematical formulation and having identified the challenges, the research questions are put forth in the next section.

3.3 Research Questions

This project is used to address the below research questions:

Q1: *How does adding additional features help to enrich existing customer segments?*

The sub questions to be answered are:

Q1a: *What would be a good machine learning model?*

When choosing a model different aspects needs to be considered such as scalability of the model and does the model meet the business goal?

Q1b: *What are the relevant features in order to segment customers?*

This can be analysed by analysing the covariance of the feature. Adding features that are relevant to the organic customer segment is key. Features that do not add value in order to determine if a customer is interested in organic product can inhibit the machine learning model from finding relevant customers.

Q1c: *What is the usability of the machine learning model and its outcome?*

The usability of the machine learning model from a business perspective is discussed.

Q2: *How does type 1 and type 2 error error affect the model and impact enrichment of customers?*

Type 1 error, in this case is when a class 1 customer (other customers) is labelled as class 0 customer. This leads to adding of customers that are not relevant. Type 2 error is when organic customers are labelled as other customers. If there is a high type 1 error, ads would be shown to customers that are not interested in organic products. This is a lost opportunity to show them ads based on their interest.

In the next section, the approach and detailed discussion on the EM algorithm is carried out.

Chapter 4

Approach

In this chapter, the manual threshold shifting approach and the EM algorithm is discussed in detail, along with the assumptions. Based on the mathematical formulation and literature review the customers can be modelled as a Gaussian mixture model. In order to fit the Gaussian mixture, the EM algorithm is used in order to estimate the parameters of the mixture. The EM algorithm is compared with the threshold shifting approach.

4.1 Assumptions

A fundamental assumption is that the customers who are interested in organic products (organic customers) and customers who are not interested in organic products (other customers) are normally distributed with mean and covariance. The underlying assumption is that each class can be modeled by a Gaussian mixture. So given a Gaussian mixture, the mixture is comprised of two Gaussian distributions. One of the distribution corresponds to customers who are interested in organic products and the other distribution belongs to customers who are not interested in organic products. An instance X $\in \mathbb{R}^D$ belong to one of the two classes with mean μ, covariance Σ and the ratio of the points that belong to each distribution is given by w.

$$X \sim \begin{cases} \mathcal{N}(\mu_0, \Sigma_0) & \text{if } X \in Z_0, \\ \mathcal{N}(\mu_1, \Sigma_1) & \text{if } X \in Z_1. \end{cases} \tag{4.1}$$

The Probability Density Function is given by

$$\mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}-\boldsymbol{\mu})}{2}\right), \tag{4.2}$$

where $\boldsymbol{X} \in \mathbb{R}^D, \boldsymbol{\mu} \in \mathbb{R}^D$ is the mean $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ is the covariance matrix, and $|.|$ is the determinant of matrix. In the next section, an intuition about the importance of covariance is given.

4.1.1 Covariance

Covariance is the measure of linear association of two variables. For instance a positive covariance between organic products and total products indicates that as organic products increases the count of total products also increases. A negative covariance indicates that as organic products increase the count of total products decrease. Mathematically, it can be defined as,

$$\mathrm{Cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]. \tag{4.3}$$

The covariance of a random variable with itself is the variance of the random variable which is given as

$$\mathrm{Cov}(X_i, X_i) = \mathrm{E}(X_i X_i) - \mathrm{E}(X_i)\mathrm{E}(X_i) = \mathrm{E}\left(X_i^2\right) - (\mathrm{E}(X_i))^2 = \mathrm{Var}(X_i). \tag{4.4}$$

In clustering, it is important to know the shape of the cluster. Understanding the covariance between features helps to get an intuition about the shape of the cluster between the features.

The relation between covariance and correlation is as follows. The correlation is a ratio that takes values from -1 to 1. A negative correlation means the variables move in opposite direction, a correlation of 0 means no correlation and a correlation of 1 means that they are strongly correlated. The correlation is given by

$$\operatorname{Cor}(X_i, X_j) = \rho = \frac{\operatorname{Cov}(X_i, X_j)}{\operatorname{Var} X_i \operatorname{Var} X_j}. \tag{4.5}$$

In Figure 4.1, 4.2 and 4.3 the shape of the cluster when the features have a positive, negative and no correlation is visualised. In the next section, the Gaussian mixture model via the EM algorithm is explained. This is an extension of the EM algorithm described in Section 2.3.

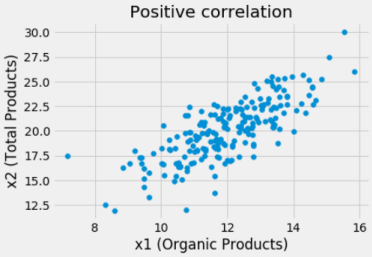

Figure 4.1: Positive correlation between features

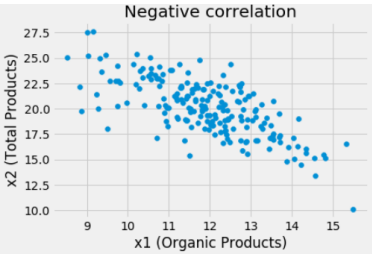

Figure 4.2: Negative correlation between features

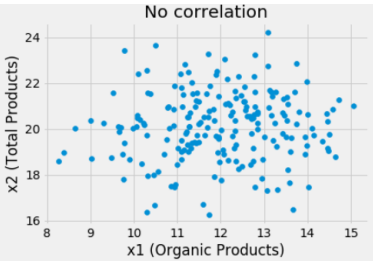

Figure 4.3: No correlation between features

4.2 Algorithm

It is assumed that the organic customers and the other customers come from two different Gaussian distributions. A Gaussian mixture model is a density model where a finite number of K (here, K=2). Gaussian distributions are combined. Given a set of observations $X = \{x_1, ..., x_N\}$, with a set of underlying ground truth labels $Z = \{z_1, ..., z_N\}$, the Gaussian mixture model tries to model the data set as a mixture of Gaussian distributions. In clustering, the K Gaussian components represent K clusters and label z_N represent the true membership of the point x_N, that is

$$\mathbf{P}\left(z_n = k\right) = w_k. \tag{4.6}$$

Without knowing the true value of the label, the joint log-likelihood of observed data is given as $X = \{x_1.....x_N\}$ as,

$$\ell(\theta|\boldsymbol{X}) = \ln \mathbf{P}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} w_k \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}. \tag{4.7}$$

The unknown parameters to be estimated are $\theta = \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k)\}_{k=1}^{K}$. In this case, K = 2.

For each observed point x_n, we obtain the membership probability to the k-th component by a straightforward application of Bayes' theorem.

$$\gamma_k\left(\boldsymbol{x}_n, \theta\right) = \mathbf{P}\left(z_n = k|\boldsymbol{x}_n, \theta\right) = \frac{w_k \mathcal{N}\left(\boldsymbol{x_n}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} w_j \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}, \tag{4.8}$$

where each $\gamma(x_n) = [\gamma_1(x_n), ..., \gamma_K(x_n)]^T \in \mathbb{R}^K$, is a normalized probability vector where each $\sum_k \gamma_k(x_n) = 1$. The data point x_n is clustered into the component for which it has the highest

membership. To fit the Gaussian Mixture Model in the EM algorithm, the Q function from (2.3) can be written as;

$$\begin{aligned} Q(\theta|\theta^{\text{old}}) &= \mathrm{E}_{Z|X,\theta^{\text{old}}}(\ell_c(\theta)) \\ &= \sum_{n=1}^{N} \mathrm{E}\left\{ \ln \left[\prod_{k=1}^{K} \left(w_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)^{I(z_n=k)} \right) \right\} \right. \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathrm{E}\left[I\left(z_n = k\right) |\boldsymbol{x}_n, \boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}} \right] \ln \left[w_k \mathcal{N}\left(\boldsymbol{x_n}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k \left(\boldsymbol{x}_n, \theta^{\text{old}} \right) \ln \left[w_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right]. \end{aligned} \tag{4.9}$$

The main purpose of the E-step in the ith iteration is to evaluate the membership probabilties based on the current estimate of the parameters by using

$$\gamma_k\left(\boldsymbol{x}_n\right) = \gamma_k\left(\boldsymbol{x}_n, \theta^{\text{old}}\right) = \frac{w_k^{\text{old}} \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}}\right)}{\sum_{j=1}^{K} w_j^{old} \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}}\right)}. \tag{4.10}$$

Given a Gaussian mixture model the goal is to maximize the likelihood function with respect to the parameters μ and Σ and mixing coefficients w_k. In this case $K = 2$, where K is the number of clusters.

1. Intitalize the means μ_k^{old}, covariances $\sum_k^{\text{old}}$ and weights w_k^{old} and evaluate the initial log likelihood.
2. **E Step** Evaluate the responsibilities using the current parameter values

$$\gamma_k\left(\boldsymbol{x}_n\right) = \frac{w_k^{\text{old}} \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}}\right)}{\sum_{j=1}^{K} w_j^{\text{old}} \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}}\right)} \tag{4.11}$$

3. **M-Step** Re-estimate parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_k\left(\boldsymbol{x}_n\right) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_k\left(\boldsymbol{x}_n\right)} \tag{4.12}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_k\left(\boldsymbol{x}_n\right) \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{old}\right)}{\sum_{n=1}^{N} \gamma_k\left(\boldsymbol{x}_n\right)} \tag{4.13}$$

$$w_k^{new} = \frac{1}{N} \sum_{n=1}^{N} \gamma_k\left(\boldsymbol{x}_n\right) \tag{4.14}$$

4. Evaluate the log likelihood and check for convergence of the log likelihood.

$$\ln \mathbf{P}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} w_k \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\} \tag{4.15}$$

4.2.1 Convergence Properties

It is possible to prove some general convergence properties of the algorithm:

- It can be shown that the EM algorithm increases monotonously the log likelihood. For proof please refer to [54].
- With a few exceptions the algorithm is not guaranteed to converge to a local maxima. Under some regularity conditions on the likelihood and on the parameter μ, $\sum$ and w ,it can be shown that the sequence of parameters converge to a local maximiser of the likelihood. Conditions that are necessary for convergence can be found in [32]. As the algorithm converges to the log likelihood, the quality of the initial estimate has an impact on the final result.

- In each iteration to evaluate K Gaussian densities for N points in the E step scales as $O\left(KND^3\right)$. The M-step requires $O\left(KND^2\right)$ to estimate the Gaussian parameters [55].

The algorithm ensures a soft clustering as it outputs probabilities of belonging to either the organic or the other customer segment. The EM algorithm is sensitive to the initial starting values [56]. The EM for GMM has close resemblance to K-means. The K-means does not estimate the covariances of the clusters but only the cluster means.

4.3 Threshold Shifting

Based on business rules, as seen in Equation (1.1), customers are segmented as part of the organic segment if the first feature, organic products are greater than a mean μ_0. In threshold shifting approach, this μ_0 of the first feature is reduced and customers are labelled as organic or other. For instance, if based on business rules the mean of the first feature, $\mu_0 >= 12$, then the customers are labelled as organic or other. In threshold shifting the μ_0 of the first feature, is reduced to for instance, $\mu_0 \geq 8$ and customers are labelled as organic or other.

Threshold shifting uses the first feature (organic products), based on business rules in order to enrich the segment. Additional features are added and the EM algorithm is applied in order to enrich the segment. Both the methods are carried out and the enrichment of the segment is observed. In the next section, numerical experiments are set up on synthetic data set and both the threshold shifting and the EM algorithm are compared.

Chapter 5

Numerical Experiments

It is proposed that adding additional features help to enrich existing customer segments. Therefore, numerical experiments are set up in order to observe the enrichment of customers when additional features are added. In this chapter, the method in which data is generated and evaluated for the numerical experiments are described. Based on this, numerical experiments are performed. The EM algorithm and the threshold shifting approaches are compared and observations are highlighted.

5.1 Data Generation

As per Section 4.1, it is assumed that customers are generated from a Gaussian mixture, with organic customers having parameters mean (μ_0), covariance (Σ_0) and weight (w_0) and the other customers with parameters mean (μ_1), covariance (Σ_1) and weight (w_1).

The use case used in the synthetic data set is organic products and the results can be generalised for other use cases. N random variables are observed, here $N = 550$, where $N_0 = 200$ and $N_1 = 350$. The data is generated with six features, that is $D = 6$. The first feature that is used based on business rules are unique organic products. The first three features pertain to the organic products bought, total products, total amount spent in a transaction (revenue), respectively and the last three features pertain to other frequently bought products. Data is generated in order to observe the impact of enrichment of customers based on different covariance structures, the covariance between features and overlapping clusters.

5.2 Setup

Several experiments are carried out in order to analyse the effectiveness of the algorithm. Five cases with varying parameters are set up based on the covariance structure, means and overlapping clusters.

In Case 1, the mean of the first feature (organic products), in the two distributions, is far apart and the clusters are linearly separable. In Case 2, the data is set up such that the mean of the first feature is not that far from each other and the clusters are not linearly separable. In Case 3, the impact of features with negative covariance to the organic product feature, in the organic distribution, is analysed. In Case 4, the distributions with different means and unitary covariance, that is when the features are not correlated with each other is analysed. In Case 5, when both the distributions have the same mean, the effectiveness of the EM algorithm is analysed.

First the approach of threshold shifting is carried out. Furthermore additional features are added one at a time. The EM algorithm is run and the impact on enrichment by adding each feature is analysed. The parameters of the organic and other distribution is not known with the AH data set.

Initialization of the parameters for organic distribution (μ_0, Σ_0, w_0) is done based on business rules and for the other distribution (μ_1, Σ_1, w_1), the rest of the data points are considered for initialization. The EM algorithm is sensitive to initialization, thus, in cases 3, 4 and 5 random initialization and random initialization with weights is compared with initialization based on business rules. In random initialization, a set of random data points are considered as the parameters of the distribution.

5.3 Evaluation

The different cases that are generated are evaluated. The customers generated from the organic distribution are labelled as 0, the customers generated from the other distribution are labelled as 1. This is the underlying ground truth in the data set.

First, results for threshold shifting approach is noted. Second, the EM algorithm by adding additional features one at a time is noted for different initialization. The output is represented as a confusion matrix as described in Section 3.1 and recall and precision is analysed.

5.4 Experiments

5.4.1 Case 1

The mean and covariance of the two distributions used to generate the data is given in (5.1). The features, pertain to the organic products bought, total products, total amount spent in a transaction (revenue), respectively and the last three features pertain to other frequently bought products. The mean of organic products vary significantly with the organic distribution and other distribution. In the organic distribution, the organic products have a strong covariance with total products and other features. In the other distribution, the covariance between other frequently bought products and organic products does not necessarily need to increase with organic products. So, the fourth feature is generated with a negative covariance. This is the intuition based on which the data set is generated. The covariance between the organic and other distribution has been generated such that the covariance between the distributions do not significantly differ. This is because with the AH data set, there can be features that do not clearly differentiate between organic and other distribution. So, the covariance of features of both the distributions can be similar.

$$\mu_0 = \left[12, 20, 40, 5, 3, 7\right], \mu_1 = \left[6, 20, 40, 5, 3, 7\right]$$

$$\Sigma_0 = \begin{bmatrix} 2 & 3 & 1 & 1 & 1 & 1 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2 & 3 & 1 & -0.1 & 1 & 1 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 2 & 1 & 1 & 1 \\ -0.1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 \end{bmatrix} \tag{5.1}$$

Graphically the plot of organic products can be seen below. In Figure 5.1 and 5.2, the customers in yellow are other customers (TN), the customers in purple are the organic customers (TP) which are found based on business rules. The blue data points (FN) are the customers that need to be found in order to enrich the segment. The blue dots are also customers generated from the organic distribution but are not included in the organic segment based on business rules. The blue dots are the customers that need to be targeted for enrichment. It can also be observed that the data points are not separable when only the first feature (organic products) are present, but when the first two features are visualized, they are separable.

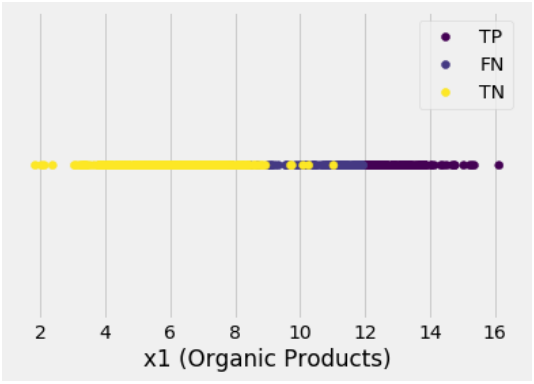

Figure 5.1: Plot of first feature

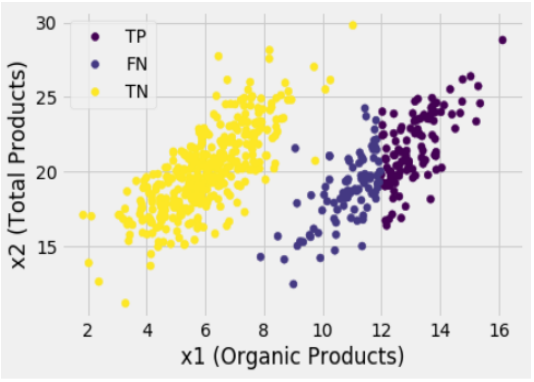

Figure 5.2: Plot of first two features

Out of the 550 customers based on business rules, 95 customers are labelled as organic customers. The 105 customers from the business rules are the customers that need to be targeted. Based on business rules, the threshold of the mean (μ_0) of the first feature (organic products) is reduced for instance from 12 to 8. From Figure 5.1 it can be seen that the yellow points (TP) are also included in the segment and the customers are not relevant. The result of the manual thresholding can be seen in Table 5.1.

In order to add only the relevant customers, customers with the blue data points, additional features are added one at a time and the EM algorithm is applied. The mean (μ_0) and covariance ($\sum_0$) of the organic distribution is initialized as the purple dots. The mean (μ_1) and covariance ($\sum_1$) of other customers is initialised by finding the mean and covariance of the yellow and blue dots. The weights, w_0 and w_1 are initialized, with w_0 as the ratio of the purple dots to all the data points and w_1 as the ratio of the green and yellow points. The business rule is used as a starting point to initialise. This is done because with the AH data set, the only parameters known about organic customers is based on the business rules. The mean and covariance with which the parameters are initialised is given in (5.2).

$$\mu_{0(\text{init})} = \left[11.98, 20, 40.05, 4.98, 3.16, 6.87\right], \mu_{1(\text{init})} = \left[6.06, 20.27, 40.02, 5.18, 3.09, 7.11\right]$$

$$\Sigma_{0(\text{init})} = \begin{bmatrix} 2.02 & 3.10 & 1.35 & 0.89 & 1.17 & 1.14 \\ 3.10 & 8.35 & 2.78 & 0.91 & 1.50 & 1.53 \\ 1.35 & 2.78 & 3.17 & 0.88 & 0.72 & 0.99 \\ 0.89 & 0.91 & 0.88 & 2.90 & 0.70 & 0.98 \\ 1.17 & 1.50 & 0.72 & 0.70 & 2.60 & 0.58 \\ 1.14 & 1.53 & 0.99 & 0.98 & 0.58 & 3.84 \end{bmatrix}, \Sigma_{1(\text{init})} = \begin{bmatrix} 1.97 & 2.91 & 0.78 & -0.24 & 0.78 & 0.90 \\ 2.91 & 7.92 & 1.87 & 0.91 & 0.62 & 0.85 \\ 0.78 & 1.87 & 1.75 & 0.92 & 0.80 & 0.71 \\ -0.24 & 0.91 & 0.92 & 2.94 & 0.74 & 0.86 \\ 0.78 & 0.62 & 0.80 & 0.74 & 2.67 & 0.89 \\ 0.90 & 0.85 & 0.71 & 0.86 & 0.89 & 3.87 \end{bmatrix} \tag{5.2}$$

First, the EM algorithm is applied with the first two features. The output of the parameters from the EM algorithm with the two features is given.

$$\mu_0 = \left[11.89, 19.67\right], \mu_1 = \left[6.05, 20.16\right]$$

$$\Sigma_0 = \begin{bmatrix} 2.04, 3.19 \\ 3.19, 7.76 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2.02, 2.95 \\ 2.95, 7.93 \end{bmatrix} \tag{5.3}$$

Similarly the third, fourth, fifth and the sixth features are added one at a time. With the EM algorithm, as seen in Table 5.1 one can observe that the organic and other customers are labelled

as required and all the customers are found. The output of the parameters of the EM algorithm with all 6 features is given in (5.4).

$$\mu_0 = \begin{bmatrix}12.07, 20.09, 40.11, 5.21, 3.34, 7.05\end{bmatrix}, \mu_1 = \begin{bmatrix}5.94, 19.80, 39.90, 4.90, 3.01, 6.94\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 2.18 & 3.23 & 1.25 & 1.27 & 1.41 & 1.11 \\ 3.23 & 8.30 & 2.82 & 1.22 & 1.88 & 1.77 \\ 1.25 & 2.82 & 3.32 & 0.96 & 1.50 & 0.99 \\ 1.27 & 1.22 & 0.96 & 3.19 & 1.43 & 1.03 \\ 1.41 & 1.88 & 1.50 & 1.43 & 3.19 & 1.23 \\ 1.11 & 1.77 & 0.99 & 1.03 & 1.23 & 4.47 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 2.17 & 3.23 & 1.13 & -0.08 & 1.17 & 1.19 \\ 3.23 & 8.50 & 2.24 & 0.98 & 1.11 & 0.81 \\ 1.13 & 2.24 & 1.94 & 1.03 & 1.08 & 0.92 \\ -0.08 & 0.98 & 1.03 & 2.85 & 0.98 & 0.57 \\ 1.17 & 1.11 & 1.08 & 0.98 & 3.00 & 1.33 \\ 1.19 & 0.81 & 0.92 & 0.57 & 1.33 & 4.27 \end{bmatrix} \tag{5.4}$$

In Table 5.1, the output of threshold shifting and the EM algorithm is given. When the first to six features are added one at a time, the results obtained are the same.

			Organic (0)		Other (1)			
Features	Method	Initialization	TP	FN	TN	FP	Recall	Precision
1	Threshold shifting	μ0 >= 8	200	0	330	20	1.00	0.91
1-6 features	EM	business rules	200	0	350	0	1.00	1.00

Table 5.1: Evaluation of threshold shifting and EM algorithm

Observations

Data set was generated with the mean of the first feature far apart from both the distributions. Manually decreasing the threshold of the mean of the first feature increases type 1 error. Additional features are added one at a time and the EM algorithm is applied. As the data points are separable, it can be observed that the algorithm outputs the same results when all the features are added. The algorithm is able to capture all the customers interested in organic products. In the next case, the algorithm is analysed with different mean and similar covariance structure as in case 1.

5.4.2 Case 2

In this case, the mean of the first feature (organic products) between the distributions is not that far apart. The covariance of both the distributions are generated with the same intuition as in Case 1. The synthetic data set is generated with parameters as mentioned in (5.5).

$$\mu_0 = \begin{bmatrix}12, 20, 40, 5, 3, 7\end{bmatrix}, \mu_1 = \begin{bmatrix}9, 20, 40, 5, 3, 7\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 2 & 3 & 1 & 1 & 1 & 1 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2 & 3 & 1 & -0.1 & 1 & 1 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 2 & 1 & 1 & 1 \\ -0.1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 \end{bmatrix} \tag{5.5}$$

In Figure 5.3 and 5.4, the data is visualised with the first feature (organic products) and with the first two features. The customers in yellow (TN) are the customers generated from the other

distribution. The customers in purple (TP) are the organic customers, which are found based on business rules. The blue data points (FN) are the customers that need to be found in order to enrich the segment. The blue dots are also customers generated from the organic distribution but are not included in the segment based on business rules. The green dots (FP) are the customers that are part of the organic segment based on business rules. The green dots are generated from the other distribution and they are the customers that are not relevant.

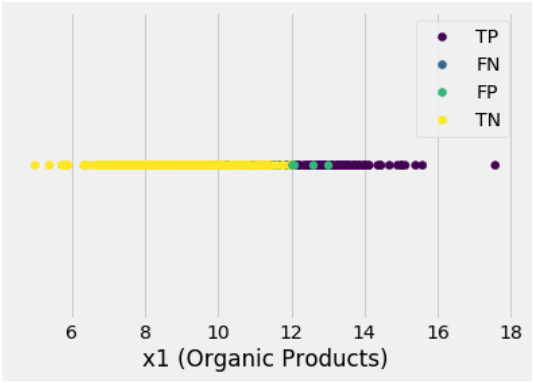

Figure 5.3: Plot of first feature

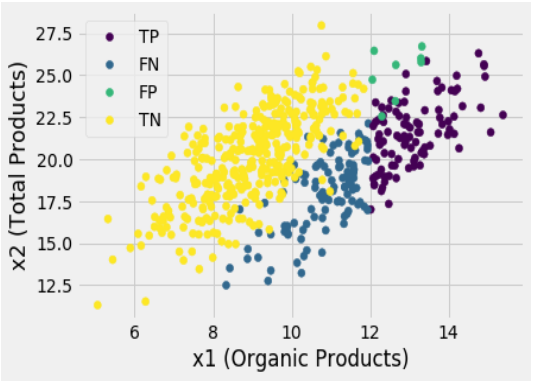

Figure 5.4: Plot of first two features

Based on business rules, 95 customers are labelled as organic customers. The customers that need to be targeted are 105 customers, that is these are the customers generated from organic distribution but are not part of the segment based on business rules. So, in total over 200 customers need to be labelled as organic customers. Manually decreasing the threshold of the mean (μ_0) of the first feature from 12 to 8 leads to an increase in type 1 error. This can be seen in Table 5.2. The precision and recall is shown in 5.6.

The parameters of the organic and other distribution is not known with the AH data set. So, the initialization for the EM aglorithm is done based on business rules. The mean μ_0 and covariance $\sum_0$ is initialized based on business rules. The mean μ_1 and covariance $\sum_1$ of other customers is initialised by finding the mean and covariance of the yellow and blue dots. The weights, w_0 and w_1 are initialized, with w_0 as the ratio of the purple dots and green dots to all the data points and w_1 as the ratio of the green and yellow points. The mean and covariance used to initialize the parameters is given in (5.6).

$$\mu_{0(\text{init})} = \begin{bmatrix}11.92, 19.59, 39.72, 5.15, 2.95, 6.97\end{bmatrix}, \mu_{1(\text{init})} = \begin{bmatrix}9.13, 20.19, 39.99, 4.77, 2.91, 6.97\end{bmatrix}$$

$$\Sigma_{0(\text{init})} = \begin{bmatrix} 2.09 & 3.36 & 1.25 & 1.08 & 1.01 & 1.41 \\ 3.36 & 7.96 & 2.24 & 1.49 & 1.04 & 1.63 \\ 1.25 & 2.24 & 3.04 & 1.50 & 0.88 & 1.28 \\ 1.08 & 1.49 & 1.50 & 3.24 & 0.70 & 1.54 \\ 1.01 & 1.04 & 0.88 & 0.70 & 2.79 & 1.11 \\ 1.41 & 1.63 & 1.28 & 1.54 & 1.11 & 3.86 \end{bmatrix}, \Sigma_{1(\text{init})} = \begin{bmatrix} 2.17 & 2.95 & 0.99 & -0.13 & 1.10 & 0.86 \\ 2.95 & 7.78 & 1.76 & 0.86 & 0.89 & 0.46 \\ 0.99 & 1.76 & 1.82 & 0.80 & 0.96 & 0.99 \\ -0.13 & 0.86 & 0.80 & 2.72 & 0.90 & 0.74 \\ 1.10 & 0.89 & 0.96 & 0.90 & 2.89 & 0.92 \\ 0.86 & 0.46 & 0.99 & 0.74 & 0.92 & 3.50 \end{bmatrix} \tag{5.6}$$

The EM algorithm is first applied with 2 features. The output of the parameters of the two distributions from the EM algorithm is given in (5.7).

$$\mu_0 = \begin{bmatrix}12.11, 20.00\end{bmatrix}, \mu_1 = \begin{bmatrix}8.99, 19.79\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix}2.30, 3.65 \\ 3.65, 9.19\end{bmatrix}, \Sigma_1 = \begin{bmatrix}1.95, 2.49 \\ 2.49, 6.10\end{bmatrix} \tag{5.7}$$

Similarly, each feature is added one at a time and the parameters are initialized as in (5.6). The impact of adding features is observed. The output of the parameters of the EM algorithm with six features is seen in (5.8).

$$\mu_0 = \left[11.93, 19.58, 39.72, 5.17, 2.95, 6.84\right], \mu_1 = \left[9.14, 20.19, 39.99, 4.77, 2.91, 6.97\right]$$

$$\Sigma_0 = \begin{bmatrix} 2.09 & 3.37 & 1.25 & 1.08 & 1.01 & 1.41 \\ 3.37 & 7.97 & 2.24 & 1.50 & 1.03 & 1.64 \\ 1.25 & 2.24 & 3.04 & 1.51 & 0.88 & 1.28 \\ 1.08 & 1.50 & 1.51 & 3.22 & 0.70 & 1.55 \\ 1.01 & 1.03 & 0.88 & 0.70 & 2.79 & 1.11 \\ 1.41 & 1.64 & 1.28 & 1.55 & 1.11 & 3.83 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2.19 & 2.95 & 0.99 & -0.13 & 1.10 & 0.86 \\ 2.95 & 7.78 & 1.76 & 0.86 & 0.89 & 0.46 \\ 0.99 & 1.76 & 1..82 & 0.80 & 0.96 & 0.99 \\ -0.13 & 0.86 & 0.80 & 2.72 & 0.90 & 0.74 \\ 1.10 & 0.89 & 0.96 & 0.90 & 2.89 & 0.92 \\ 0.86 & 0.46 & 0.99 & 0.74 & 0.92 & 3.50 \end{bmatrix} \tag{5.8}$$

The results obtained when threshold shifting is applied to the first feature (organic products) and when the EM algorithm is applied to features one at a time, is seen in Table 5.2.The output of the EM algorithm with all 6 features is visualised in 5.5 and precision and recall is given in 5.6.

			Organic (0)		Other (1)			
Features	Method	Initialization	TP	FN	TN	FP	Recall	Precision
1	Threshold shifting	μ0 >= 8	200	0	84	266	1.00	0.43
2	EM	business rules	191	9	333	17	0.96	0.92
3	EM	business rules	193	7	333	17	0.97	0.92
5	EM	business rules	192	8	341	9	0.96	0.96
5 features without 4th feature	EM	business rules	195	5	339	11	0.98	0.95
6	EM	business rules	195	5	342	8	0.98	0.96

Table 5.2: Evaluation of threshold shifting and EM algorithm

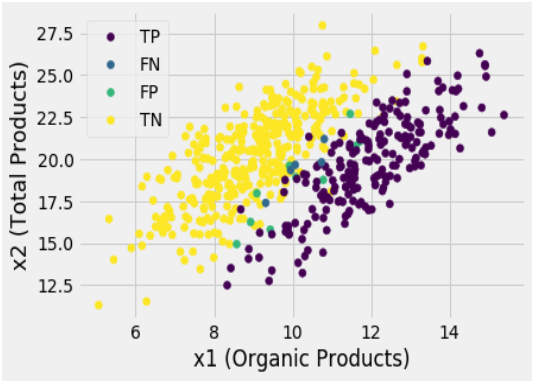

Figure 5.5: Output of EM algorithm with six features

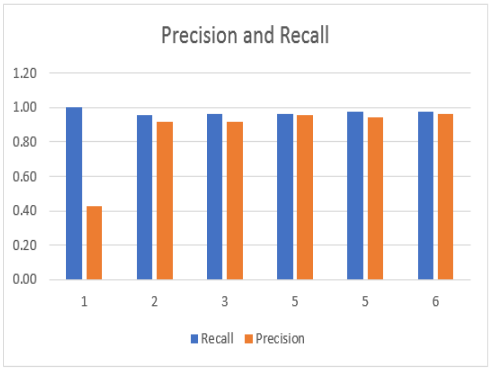

Figure 5.6: Precision and Recall

Observations

In this case, the data set for the organic and other distribution was generated such that the yellow and the blue data points overlap as seen in Figure 5.4. So, there is no clear separation between organic and other customers. Based on business rules, manually reducing the threshold of the first feature to enrich the segment leads to also adding customers from the other segment. The yellow data points also get added and hence such an approach cannot be used. The EM algorithm is applied on the first two features, the algorithm outputs good results and is able to find most of the purple and the blue dots. But it also captures some of the yellow data points. So, in order to add only relevant customers

to the segment and decrease type 1 error and type 2 error, additional features are added one at a time. It can be observed that adding additional features helps to decrease type 1 and type 2 error. The algorithm is also applied on five features, without the fourth feature. The fourth feature has a negative covariance to organic products in the the other distribution. Removing the fourth feature, increases type 1 error but decreases type 2 error. Compared to Case 1, where there is no type 1 and type 2 error, the type 1 and type 2 error is higher in this case.

5.4.3 Case 3

The mean of both distributions are the same as in Case 2. In this case, the impact on enrichment when negatively correlated features to the first feature (organic products) are added in the organic distribution is analysed. The last two features that are generated in the organic distribution are negatively correlated with respect to the first feature. The first feature is the feature used in the business rules. The covariance of rest of the features of both the distributions are generated with the same intuition as in Case 1. The parameters of the generated data is given in (5.9).

$$\mu_0 = \left[12, 20, 40, 5, 3, 7\right], \mu_1 = \left[9, 20, 40, 5, 3, 7\right]$$

$$\Sigma_0 = \begin{bmatrix} 2 & 3 & 1 & 1 & -0.1 & -3 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ -0.1 & 1 & 1 & 1 & 3 & 1 \\ -3 & 1 & 1 & 1 & 1 & 4 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2 & 3 & 1 & -0.1 & 1 & 1 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 2 & 1 & 1 & 1 \\ -0.1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 \end{bmatrix} \tag{5.9}$$

The data is visualised with the first two features in Figure 5.7 and in 5.8. The customers in yellow (TN) are other customers, the customers in purple (TP) are the organic customers, which are found based on business rules. The blue data points (FN) are the customers that need to be found in order to enrich the segment. The green data points (FP) are the customer that are part of the organic segment based on business rules. But they are generated from the other distribution and are not relevant. It can be observed that the data points are not lineally separable.

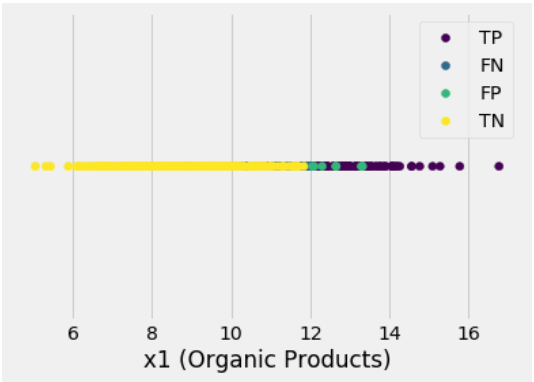

Figure 5.7: Plot of first feature

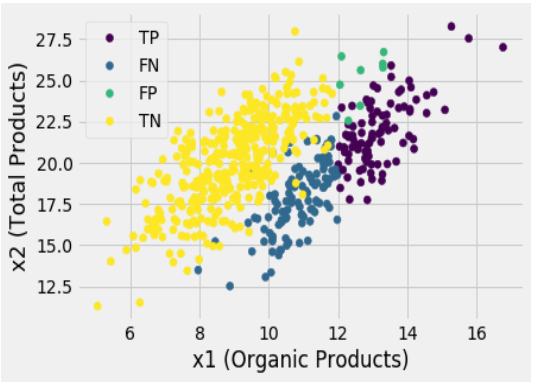

Figure 5.8: Plot of first two features

Manually reducing the threshold of the mean of the first feature (μ_0) gives a high type 1 error. The results can be seen in Table 5.3. The precision and recall is visualised in 5.10.

Additional features are added one at a time and the EM algorithm is applied. With the AH data set, the mean and covariance of the organic and other distribution is not known. So, the organic segment is initialized based on business rules. The mean (μ_0) and the covariance (Σ_0) for the EM algorithm are initialised with the organic distribution as the data points from the business rules. The

mean (μ_1) and covariance (Σ_1) of other customers is initialised by finding the mean and covariance of the yellow and blue dots. The initialization of the parameters is given in (5.10).

$$\mu_{0(\text{init})} = \begin{bmatrix}13.90, 21.80, 40.02, 5.01, 2.99, 6.72\end{bmatrix}, \mu_{1(\text{init})} = \begin{bmatrix}9.39, 19.79, 39.93, 4.89, 2.92, 7.09\end{bmatrix}$$

$$\Sigma_{0(\text{init})} = \begin{bmatrix} 0.49 & 0.75 & 0.42 & 0.20 & -0.09 & 0.04 \\ 0.75 & 4.19 & 1.02 & -0.67 & 0.87 & 2.05 \\ 0.42 & 1.02 & 2.74 & 0.90 & 1.21 & 1.80 \\ 0.20 & -0.67 & 0.90 & 3.18 & 1.25 & 1.11 \\ -0.09 & 0.87 & 1.21 & 1.25 & 3.17 & 1.00 \\ 0.04 & 2.05 & 1.80 & 1.11 & 1.00 & 5.72 \end{bmatrix}$$

$$\Sigma_{1(\text{init})} = \begin{bmatrix} 2.37 & 1.71 & 0.52 & -0.35 & 0.77 & 0.66 \\ 1.71 & 6.98 & 1.66 & 1.03 & 0.89 & 0.0.86 \\ 0.52 & 1.66 & 2.08 & 0.89 & 0.89 & 0.77 \\ -0.35 & 1.03 & 0.89 & 2.97 & 0.96 & 0.80 \\ 0.77 & 0.89 & 0.89 & 0.96 & 2.72 & 0.75 \\ 0.66 & 0.86 & 0.77 & 0.80 & 0.75 & 3.91 \end{bmatrix} \tag{5.10}$$

First the EM algorithm is applied on the first two features. The output of the parameters of the EM algorithm with two features is given in 5.11.

$$\mu_0 = \begin{bmatrix}12.13, 19.82\end{bmatrix}, \mu_1 = \begin{bmatrix}9.28, 20.06\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix}1.89, 3.55 \\ 3.55, 8.82\end{bmatrix}, \Sigma_1 = \begin{bmatrix}2.36, 2.71 \\ 2.71, 7.83\end{bmatrix} \tag{5.11}$$

Similarly, each feature is added one at a time and the EM algorithm is applied. The output of the parameters of the EM algorithm with six features is given in (5.12).

$$\mu_0 = \begin{bmatrix}11.90, 19.80, 40.02, 5.01, 2.99, 7.28\end{bmatrix}, \mu_1 = \begin{bmatrix}9.11, 20.41, 40.07, 5.01, 2.92, 6.91\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 2.01 & 3.38 & 1.22 & 1.03 & -0.14 & -1.01 \\ 3.38 & 8.63 & 1.96 & 0.60 & 0.52 & -0.25 \\ 1.22 & 1.96 & 2.87 & 1.03 & 0.58 & 0.20 \\ 1.03 & 0.60 & 1.03 & 2.89 & 0.66 & 0.06 \\ -0.14 & 0.52 & 0.58 & 0.66 & 2.30 & 0.85 \\ -1.01 & -0.25 & 0.20 & 0.06 & 0.85 & 4.03 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 2.19 & 2.94 & 1.12 & -0.24 & 1.28 & 1.15 \\ 2.94 & 7.29 & 2 & 0.81 & 1.11 & 0.98 \\ 1.12 & 2 & 2.05 & 0.99 & 1.22 & 0.99 \\ -0.24 & 0.81 & 0.99 & 3.10 & 1.08 & 0.68 \\ 1.28 & 1.11 & 1.22 & 1.08 & 3.37 & 1.12 \\ 1.15 & 0.98 & 0.99 & 1 & 1 & 4 \end{bmatrix} \tag{5.12}$$

The results obtained when threshold shifting is applied to the first feature (organic products) and when the EM algorithm is applied to features one at a time, is seen in Table 5.3. Initialization of the EM algorithm based on business rules, random initialization and random initialization with weights is carried out. Precision and recall is given in 5.10 and the output of the EM algorithm with all 6 features is visualised in 5.9.

			Organic (0)		Other (1)			
Features	Method	Initialization	TP	FP	TN	FP	Recall	Precision
1	Threshold shifting	μ0 >= 10	188	12	266	84	0.94	0.69
2	EM	business rules	183	17	341	9	0.92	0.95
2	EM	random initialization	118	82	155	195	0.59	0.38
2	EM	w_0 = 0.40, w_2 =0.60	112	88	170	180	0.56	0.38
3	EM	business rules	193	7	333	17	0.97	0.92
4	EM	business rules	188	12	339	11	0.94	0.94
5	EM	business rules	196	4	337	13	0.98	0.94
5	EM	random initialization	198	2	334	15	0.99	0.93
5	EM	w_0 = 0.70, w_2 = 0.30	197	3	335	15	0.99	0.93
6	EM	business rules	200	0	340	10	1.00	0.95
6	EM	random initialization	199	1	340	10	1.00	0.95
6	EM	w_0 = 0.90, w_2 = 0.10	199	1	340	10	1.00	0.95

Table 5.3: Evaluation of threshold shifting and EM algorithm with different initialization

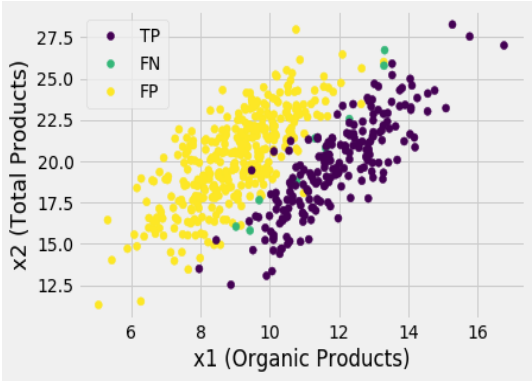

Figure 5.9: Output of EM algorithm with six features

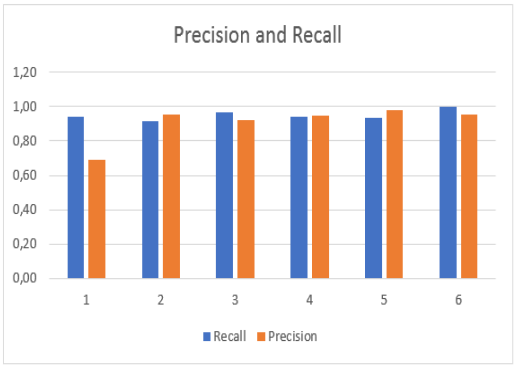

Figure 5.10: Precision and recall

Observations

In this case, the data set for the organic and other distribution was generated such that the yellow and the blue data points overlap as seen in Figure 5.7. So, there is no clear separation between organic and other customers. Features that have a negative covariance with the first feature, organic products, in the organic distribution is added. It can be observed that based on business rules, manually decreasing the threshold to enrich the segment leads to adding customers from the other segment. The yellow data points would also get added and hence such an approach cannot be used.

The EM algorithm is applied by adding features one at a time. Using the first two features, that is organic products and total products, the EM algorithm outputs good results and is able to find most of the purple and the blue dots. But it also captures some of the yellow data points. So, in order to add only relevant customers to the segment and decrease type 2 error, additional features are added. Analysis for adding additional features was done and it can be observed that adding additional features helps to decrease type 1 and type 2 error. Also, addition of negatively correlated feature to the organic segment, does not negatively impact the enrichment. When the EM algorithm is applied with six features, Case 3 gives no type 2 error compared to Case 2. This can be attributed to the fact of adding the negatively correlated features to the organic distribution. The negatively correlated features help to distinguish between organic and other distribution.

Random initialization and initialization based on weights was also done. It can be seen that random initialization and random initialization based on weights, gives a poorer performance as compared to initialization based on business rules.

5.4.4 Case 4

In Case 4, the covariances are unitary in both distributions but with different means. Unitary covariance is chosen in order to observe the behavior when the features are not correlated with one another. The data is generated with mean and covariance as in (5.13).

$$\mu_0 = \begin{bmatrix}12, 20, 40, 5, 3, 7\end{bmatrix}, \mu_1 = \begin{bmatrix}9, 20, 40, 5, 3, 7\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix}, \Sigma_1 = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\end{bmatrix} \tag{5.13}$$

The data is visualised with the first two features in Figure 5.11 and in 5.12. The customers in yellow (TN) are other customers, the customers in purple (TP) are the organic customers, which are found based on business rules. The blue data points (FN) are the customers that need to be found in order to enrich the segment. The green data points (FP) are the customer that are part of the organic segment based on business rules. But they are generated from the other distribution and are not relevant. It can be observed that the data points are not lineally separable.

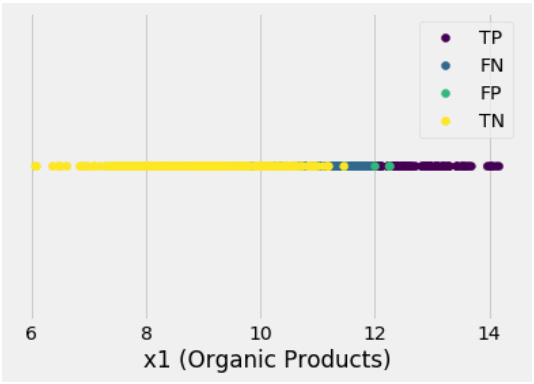

Figure 5.11: Plot of first feature

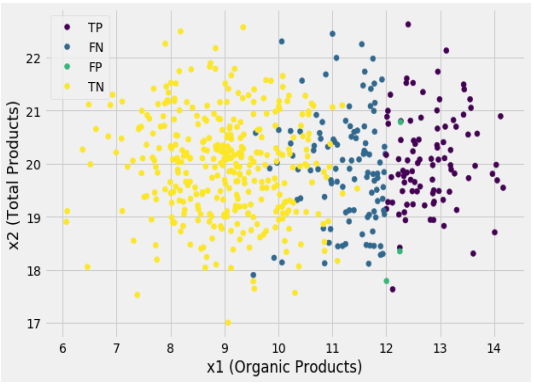

Figure 5.12: Plot of first two features

In this case, it can be noted that based on business rules, 99 customers are labelled as organic from the organic distribution. 101 customers that are part of the organic distribution, but have not been labelled as organic based on business rules. Manually reducing the threshold of the mean of the first feature (μ_0) gives a high type 1 error. The results are shown in Table 5.4.

Additional features are added one at a time and the EM algorithm is applied. The mean (μ_0) and the covariance (Σ_0) for the EM algorithm are initialised with the organic distribution as the data points from the business rules, the purple and green dots. The mean (μ_1) and covariance (Σ_1) of other customers is initialised by finding the mean and covariance of the yellow and blue dots. The initialization of the parameters for the EM algorithm is given in (5.14).

$$\mu_{0(\text{init})} = \begin{bmatrix}12.71, 20.02, 39.97, 4.99, 2.89, 7.24\end{bmatrix}, \mu_{1(\text{init})} = \begin{bmatrix}9.51, 20.03, 39.97, 4.99, 2.97, 6.98\end{bmatrix}$$

$$\Sigma_{0(\text{init})} = \begin{bmatrix}0.30 & -0.01 & -0.03 & 0.06 & 0.00 & 0.03\\ -0.01 & 0.92 & 0.05 & -0.01 & 0.07 & -0.00\\ -0.03 & 0.05 & 0.87 & -0.02 & -0.05 & 0.14\\ 0.06 & -0.01 & -0.02 & 1.02 & -0.14 & 0.06\\ 0.00 & 0.07 & -0.05 & -0.14 & 0.98 & 0.16\\ 0.03 & -0.00 & 0.14 & 0.06 & 0.16 & 1.05\end{bmatrix}$$

$$\Sigma_{1(\text{init})} = \begin{bmatrix} 1.72 & 0.00 & -0.08 & -0.00 & -0.01 & 0.03 \\ 0.00 & 0.94 & 0.04 & 0.04 & -0.03 & -0.08 \\ -0.08 & 0.04 & 1.00 & -0.03 & 0.09 & -0.05 \\ -0.00 & 0.04 & -0.03 & 0.92 & -0.03 & -0.01 \\ -0.01 & -0.03 & 0.09 & -0.03 & 1.09 & -0.09 \\ 0.03 & -0.08 & -0.05 & -0.01 & -0.09 & 0.95 \end{bmatrix} \tag{5.14}$$

The EM algorithm is applied with the first 2 features. The algorithm outputs mean and covariance as seen in (5.15).

$$\mu_0 = \begin{bmatrix} 11.90, 19.91 \end{bmatrix}, \mu_1 = \begin{bmatrix} 9.01, 20.10 \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 0.96, 0.10 \\ 0.10, 0.89 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1.09, 0.13 \\ 0.13, 0.94 \end{bmatrix} \tag{5.15}$$

Features are added one at a time. The means and covariances of both the distributions are initialized based on business rules. The effect of adding features feature is observed through the type 1 and type 2 error. The output of the mean and covariance from the EM algorithm with 6 features is given (5.16).

$$\mu_0 = \begin{bmatrix} 12.29, 19.93, 39.94, 4.98, 2.95, 7.27 \end{bmatrix}, \mu_1 = \begin{bmatrix} 9.33, 20.06, 39.98, 4.99, 2.99, 6.95 \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 0.63 & 0.06 & -0.002 & 0.05 & 0.06 & 0.02 \\ 0.06 & 0.94 & 0.13 & -0.05 & 0.09 & -0.02 \\ -0.002 & 0.13 & 0.96 & 0.003 & -0.04 & 0.10 \\ 0.05 & -0.05 & 0.00 & 0.98 & -0.18 & 0.10 \\ 0.06 & 0.09 & -0.04 & -0.18 & 0.96 & 0.18 \\ 0.02 & -0.02 & 0.10 & 0.10 & 0.18 & 0.96 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 1.56 & 0.07 & -0.08 & 0.00 & 0.01 & -0.02 \\ 0.07 & 0.92 & 0.01 & 0.05 & -0.05 & -0.07 \\ -0.08 & 0.01 & 0.97 & -0.04 & 0.10 & -0.06 \\ 0.002 & 0.05 & -0.04 & 0.92 & -0.00 & -0.03 \\ 0.01 & -0.05 & 0.10 & -0.00 & 1.10 & -0.10 \\ -0.02 & -0.07 & -0.06 & -0.03 & -0.10 & 0.96 \end{bmatrix} \tag{5.16}$$

The result of the confusion matrix is shown in Table 5.4 along with the output of the EM algorithm with 6 features in Figure 5.13. Initialization of the EM algorithm based on business rules, random initialization and random initialization with weights is carried out. Precision and recall is visualized in 5.14. In the next subsection, observations on this case is highlighted.

			Organic (0)		Other (1)			
Features	Method	Initialization	TP	FN	TN	FP	Recall	Precision
1	Threshold shifting	μ0 >= 10	196	4	290	60	0.98	0.76563
2	EM	business rules	178	22	331	19	0.89	0.90
2	EM	random initialization	188	12	320	30	0.94	0.86
2	EM	w_0 = 0.40, w_2 =0.60	190	10	311	39	0.95	0.83
3	EM	business rules	175	25	336	14	0.88	0.93
3	EM	random initialization	190	10	318	32	0.95	0.86
3	EM	w_0 = 0.70, w_2 = 0.30	190	10	318	32	0.95	0.86
4	EM	business rules	167	33	342	8	0.84	0.95
4	EM	random initialization	191	9	318	32	0.96	0.86
4	EM	w_1 = 0.70 , w_2 = 0.30	191	9	307	43	0.96	0.82
5	EM	business rules	168	32	341	9	0.84	0.95
5	EM	random initialization	191	9	318	32	0.96	0.86
5	EM	w_0 = 0.70, w_2 = 0.30	193	7	308	42	0.97	0.82
6	EM	business rules	169	31	336	14	0.85	0.92
6	EM	random initialization	159	11	319	31	0.94	0.84
6	EM	w_0 = 0.80, w_2 = 0.20	193	7	300	50	0.97	0.79

Table 5.4: Evaluation of threshold shifting and EM algorithm with different initialization

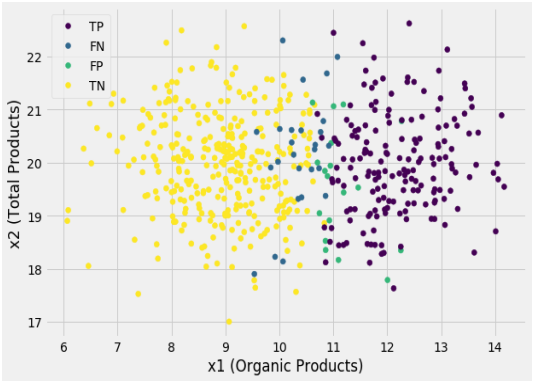

Figure 5.13: Output of EM algorithm with six features

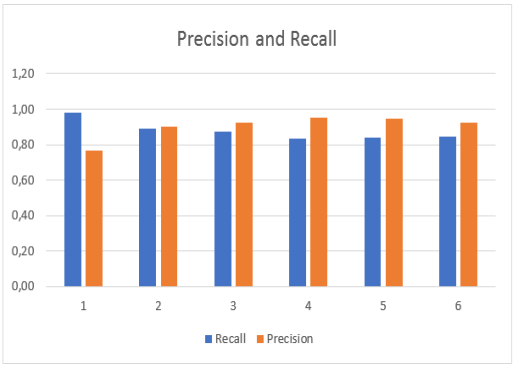

Figure 5.14: Precision and recall

Observations

Synthetic data was generated with different means and unitary covariance. The data points are not linearly separable. Manually decreasing the mean (μ_0) based on the first feature gives a very high type 1 error. In the case when the first 2 features are considered, the EM algorithm has a good precision and recall. It can be observed that random initialization and initialization of weights gives a higher type 1 error compared to initialization based on business rules. It can also be observed that random initialization with initialization of weights helps to reduce type 2 error, but increases type 1 error. As the features add no value and are not correlated to each other, the type 2 error is the least when the EM algorithm is applied on the first two features. In this case adding additional features only helps to decrease type 1 error but does not help to decrease type 2 error.

5.5 Case 5

In this case, the mean of the organic distribution and other distribution are the same but with different covariances. The data is generated with same mean in both the distributions in order to observe the effectiveness of the algorithm when the clusters are entirely overlapping. The covariance of both the distributions are generated with the same intuition as in Case 1. The data is generated with the parameters as in (5.17).

$$\mu_0 = \left[12, 20, 40, 5, 3, 7\right], \mu_1 = \left[12, 20, 40, 5, 3, 7\right]$$

$$\Sigma_0 = \begin{bmatrix} 2 & 3 & 1 & 1 & 1 & 1 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 2 & 3 & 1 & -0.1 & 1 & 1 \\ 3 & 8 & 2 & 1 & 1 & 1 \\ 1 & 2 & 2 & 1 & 1 & 1 \\ -0.1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 \end{bmatrix} \tag{5.17}$$

The data is visualised with the first two features in Figure 5.15 and in 5.16. The customers in yellow (TN) are other customers, the customers in purple (TP) are the organic customers, which are found based on business rules. The blue data points (FN) are the customers that need to be found in order to enrich the segment. The green data points (FP) are the customer that are part of the organic segment based on business rules. But they are generated from the other distribution and are not relevant. The data points are entirely overlapping.

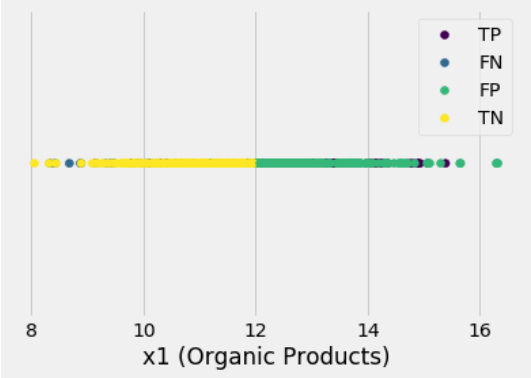

Figure 5.15: Plot of first feature

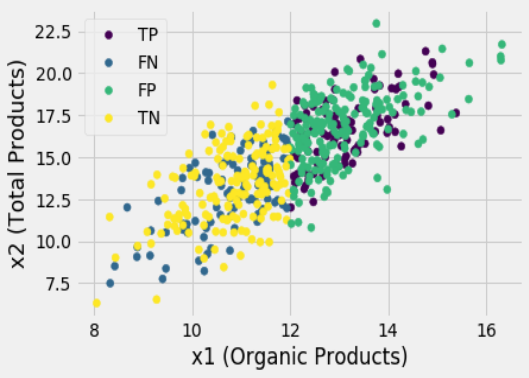

Figure 5.16: Plot of first two features

Based on business rules, 93 customers are labelled as organic from the organic segment. Over 188 customers from the other segment are labelled as organic based on business rules. The customers to target are the 107 customers in blue dots and the rest are from the other distribution. Manually reducing the threshold of the mean (μ_0) of the first feature gives a high type 1 error. The results are shown in 5.17.

Additional features are added one at a time and the EM algorithm is applied. The algorithm is initialized based on business rules. The mean (μ_0) and the covariance (Σ_0) of the organic distribution is initialised as the data points from the business rules. The mean $(\mu_1)$and covariance (Σ_1) of other customers is initialised by finding the mean and covariance of the yellow and blue dots. The initialization of the parameters for the EM algorithm is given in (5.18).

$$\mu_{0(\text{init})} = \begin{bmatrix}13.10, 21.57, 40.46, 5.75, 3.54, 7.62\end{bmatrix}, \mu_{1(\text{init})} = \begin{bmatrix}10.82, 18.17, 40.46, 5.75, 3.54, 7.62\end{bmatrix}$$

$$\Sigma_{0(\text{init})} = \begin{bmatrix} 0.68 & 0.93 & 0.39 & 0.44 & 0.40 & 0.44 \\ 093 & 3.96 & 0.98 & 0.06 & 0.07 & -0.11 \\ 0.39 & 0.98 & 2.45 & 1.05 & 0.31 & 0.95 \\ 0.44 & 0.06 & 1.05 & 3.56 & 0.17 & 1.57 \\ 0.40 & 0.07 & 0.31 & 0.17 & 2.63 & 0.97 \\ 0.44 & -0.11 & 0.95 & 1.57 & 0.97 & 3.31 \end{bmatrix}$$

$$\Sigma_{1(\text{init})} = \begin{bmatrix} 0.70 & 1.05 & 0.28 & 0.08 & 0.33 & 0.41 \\ 1.05 & 5.14 & 0.70 & 0.75 & -0.19 & 0.15 \\ 0.28 & 0.70 & 1.79 & 0.76 & 0.48 & 0.58 \\ 0.08 & 0.75 & 0.76 & 2.39 & 0.50 & 0.59 \\ 0.33 & -0.19 & 0.48 & 0.50 & 2.38 & 0.42 \\ 0.41 & 0.15 & 0.58 & 0.59 & 0.42 & 3.11 \end{bmatrix} \tag{5.18}$$

The algorithm with first 2 features outputs mean and covariance as seen in (5.19).

$$\mu_0 = \begin{bmatrix}13.03, 16.79\end{bmatrix}, \mu_1 = \begin{bmatrix}11.36, 13.88\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix}1.29, 1.53 \\ 1.53, 4.42\end{bmatrix}, \Sigma_1 = \begin{bmatrix}1.42, 2.14 \\ 2.14, 6.85\end{bmatrix} \tag{5.19}$$

Similarly, features are added one at a time and the output of the parameters of the EM algorithm with six features is given in (5.20).

$$\mu_0 = \begin{bmatrix}13.00, 21.80, 40.53, 5.22, 3.25, 7.13\end{bmatrix}, \mu_1 = \begin{bmatrix}11.25, 18.65, 39.44, 4.67, 2.68, 6.77\end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 1.23 & 1.44 & 0.65 & 0.16 & 0.88 & 0.87 \\ 1.44 & 4.32 & 1.22 & 0.25 & 0.63 & 0.23 \\ 0.65 & 1.22 & 2.17 & 1.02 & 0.96 & 1.19 \\ 0.16 & 0.25 & 1.02 & 3.38 & 1.00 & 1.30 \\ 0.88 & 0.63 & 0.96 & 1.00 & 3.14 & 1.21 \\ 0.87 & 0.23 & 1.19 & 1.30 & 1.21 & 3.99 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1.33 & 1.91 & 0.56 & 0.15 & 0.74 & 0.87 \\ 1.91 & 6.38 & 1.07 & 0.86 & 0.39 & 0.87 \\ 0.56 & 1.07 & 1.81 & 0.77 & 0.64 & 0.86 \\ 0.15 & 0.86 & 0.77 & 2.47 & 0.58 & 0.71 \\ 0.74 & 0.39 & 0.64 & 0.58 & 2.50 & 0.73 \\ 0.87 & 0.87 & 0.86 & 0.71 & 0.73 & 3.29 \end{bmatrix} \quad (5.20)$$

The results obtained when threshold shifting is applied to the first feature (organic products) and when the EM algorithm is applied to features one at a time, is seen in Table 5.17. Initialization of the EM algorithm based on business rules, random initialization and random initialization with weights is carried out. Precision and recall is given in 5.19 and the output of the EM algorithm with all 6 features is visualised in 5.18.

			Organic (0)		Other (1)			
Features	Method	Initialization	TP	FN	TN	FP	Recall	Precision
1	Threshold shifting	μ0 >= 10	189	11	17	333	0.95	0.36
2	EM	business rules	85	115	176	174	0.43	0.33
2	EM	random initialization	95	105	147	203	0.48	0.32
2	EM	w_0 = 0.40, w_2 =0.60	86	114	170	180	0.43	0.32
3	EM	business rules	74	126	186	164	0.37	0.31
3	EM	random initialization	83	117	194	156	0.42	0.35
3	EM	w_0 = 0.40, w_2 = 0.60	78	122	165	184	0.39	0.30
4	EM	business rules	71	129	178	172	0.36	0.29
4	EM	random initialization	84	116	154	196	0.42	0.30
4	EM	w_1 = 0.40 , w_2 = 0.60	75	125	171	179	0.38	0.30
5	EM	business rules	69	131	179	171	0.35	0.29
5	EM	random initialization	78	122	156	194	0.39	0.29
5	EM	w_0 = 0.70, w_2 = 0.30	94	106	135	215	0.47	0.30
6	EM	business rules	74	126	167	183	0.37	0.29
6	EM	random initialization	72	126	149	201	0.36	0.26
6	EM	w_0 = 0.80, w_2 = 0.20	95	105	114	236	0.48	0.29

Figure 5.17: Evaluation of threshold shifting and EM algorithm with different initialization

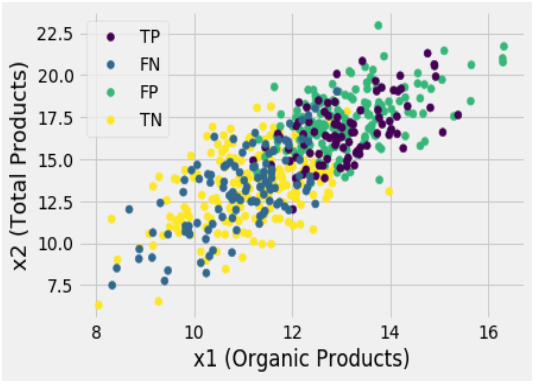

Figure 5.18: Output of EM algorithm with six features

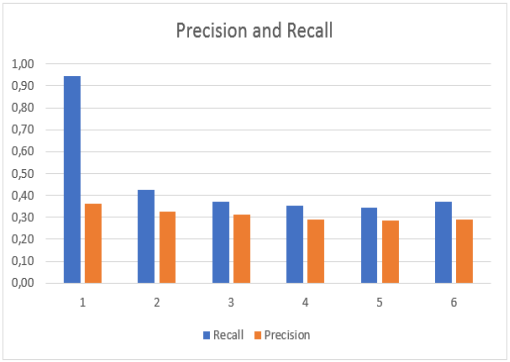

Figure 5.19: Precision and recall

Observations

Data set with two distributions having the same mean and different covariance was generated. This data was generated in order to analyse the impact of the algorithm when dealing with overlapping clusters. Based on business rules, manually decreasing the mean (μ_0) of the first feature gives a very high type 1 error. Additional features are added and the EM algorithm is applied. The algorithm fails to find the customers that needs to be enriched. When additional features are added one at a time,

it can be observed that the performance of the algorithm only worsens. So, in this case, when the clusters are entirely overlapping with different covariance, the algorithm fails to find the data points. It can also be observed that the more the clusters overlap, the EM algorithm takes longer to converge. With random initialization and with initialization of weights, the algorithm has a very high type 1 error.

5.6 Observations

Synthetic data set was generated and the effect of covariance, adding additional features and different initialization was observed. In Cases 1, 2 and 3 it can be observed that adding of additional features does help to capture the customers by decreasing type 1 and type 2 error. It can also be seen that performance is favorable with initialization based on business rules than with random or initialization based on weights. When initialized based on weights, the type 1 error increases.

In Cases 1, 2 and 3 it can be seen that adding of features does help to enrich the segment. In Case 1, when the means of both the distributions are far apart, the algorithm with six feature is able to capture all the customers in the organic segment. In Case 2, when the means of the clusters are not that far apart and the clusters are overlapping, there is a type 1 and type 2 error, compared to Case 1, when the algorithm is applied with six features. In Case 3, with negative correlated features to the organic product feature in the organic distribution, the algorithm, when applied with six features, is able to capture all the customers. In Case 4, when the features have no relation with each other, adding of additional features does not give a better performance as the means of the features are the same and the features do not add value. Although adding of the features helps to decrease type 1 error, the type 2 error increases. In Case 5, when the clusters are overlapping, the EM algorithm is not able to capture all the customers that need to be targeted. It can also be observed the as the overlapping of the clusters increase, it takes longer for the EM algorithm to converge.

Chapter 6

Application to AH Data Set

Based on the approach carried out with the synthetic data set, a similar approach is applied to the AH data set. With the synthetic data set, the underlying ground truth was established based on the assumption that customers are drawn from organic and other distribution. So, customers drawn from the organic distribution are labelled as 0 and customers drawn from the other distribution are labelled as 1. This is the underlying ground truth in the synthetic data set. With the synthetic data set, the impact of adding additional features one at a time was analysed based on type 1 and type 2 error. In order to do the same evaluation with the AH data set, an underlying ground truth needs to be established. With the AH data set the underlying ground truth is not known. First, a ground truth is established in order to evaluate the model. Then, the results are analysed.

The use case selected for the AH data set is organic products. As described in the mathematical formulation in Chapter 3, there is a feature space consisting of the features from business rules and other additional features. In total, there are thirty one features. The feature used from the business rules is count of unique organic products bought (organic products). The other features that are considered are features such as organic products, total products, average spend per week and other frequently bought products.

6.1 Evaluation

With the AH data set, the underlying ground truth is not known. In order to establish the underlying ground truth and carry out an offline evaluation, initially the data set is clustered with over thirty one features. The EM algorithm is used to cluster this data set with random initialization. It is assumed that there are 6 components in the data set. Out of the 6 clusters, the cluster with the highest mean of organic products and the second highest mean of organic products are the customers that are organic customers. A ground truth is established so that it can be shown that adding additional features, apart from the features used in the business rules, helps to enrich the segment. It is observed that out of 65,014 customers, over 21,100 customers need to be labelled as organic customers.

The data is visualised with the first and second feature in Figure 6.1 and 6.2. The purple dots are the customers labelled as organic. It can be seen that based on the features selected, the clusters are overlapping and not linearly separable.

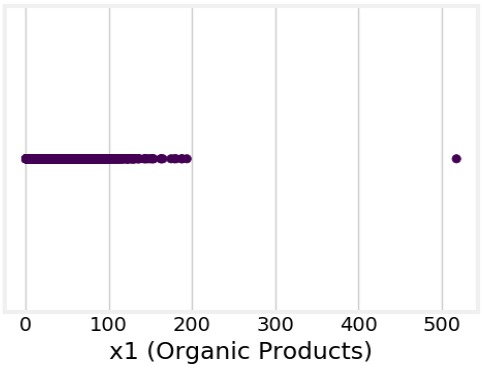

Figure 6.1: Plot of first feature

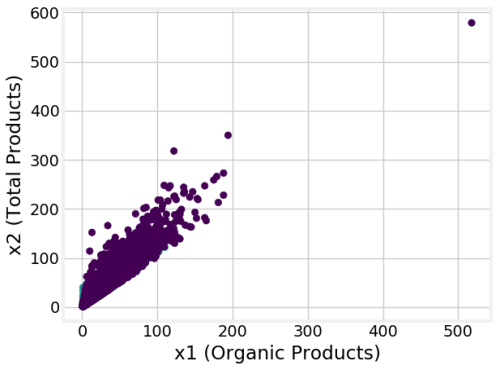

Figure 6.2: Plot of first two features

Based on business rules the mean of the first feature, organic products is 8. Customers that have bought organic products that is greater than 8 are labelled as organic. The mean of this threshold is decreased with the first feature. The results are seen in Table 6.1. Precision and recall is seen in Figure 6.3. Manually, decreasing the threshold with using the first feature, organic products as per business rules, gives a high type 1 error.

An additional feature, count of total products bought (total products), is added and the EM algorithm is applied. Similarly, additional features that are highly correlated to the organic segment, based on business rules are added one at a time. It can be seen that as top correlated features to the organic segment are added one at a time, the type 1 and type 2 error decreases. With nine features, which are the top correlated features to the organic segment, the type 1 and type 2 error decreases. Randomly adding features after the top nine correlated features, it can be observed that the type 2 error increases.

The output of the EM algorithm and threshold shifting is seen in Table 6.1. Precision and recall is seen in 6.3.

			Organic (0)		Other (1)			
		Total	21100		43914			
Features	Method	Initialization	TP	FN	TN	FP	Recall	Precision
1	Threshold shifting	μ0 >= 6	14438	6662	32609	11305	0.68	0.56
2	EM	business rules	13582	7518	37371	6543	0.64	0.67
3	EM	business rules	14320	6780	36109	7805	0.68	0.65
5	EM	business rules	14474	6626	35999	7915	0.69	0.65
9	EM	business rules	15339	5761	36179	7735	0.73	0.66
10	EM	business rules	15397	5703	36086	7828	0.73	0.66
11	EM	business rules	15850	5250	35678	8236	0.75	0.66
15	EM	business rules	17437	4304	32514	11400	0.80	0.60
20	EM	business rules	18570	2530	31242	12674	0.88	0.59
25	EM	business rules	20661	439	29653	14261	0.98	0.59
31	EM	business rules	20686	414	28907	15007	0.98	0.58

Table 6.1: Evaluation of threshold shifting and EM algorithm

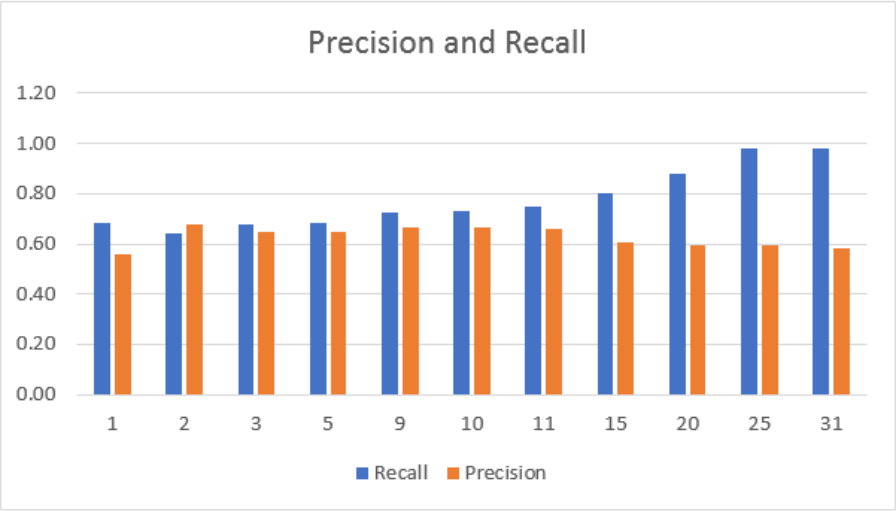

Figure 6.3: Precision and recall

6.2 Observations

With the AH data set, it can be seen that, with the available thirty one features, the data points are close together and not linearly separable. There are not distinct features that clearly distinguish organic and the other segment. It can be observed that adding highly correlated features to the organic segment help to enrich the segment. Adding additional features that have a low correlation to the organic segment, increases the type 1 and type 2 errors. It can be seen that as highly correlated features are added one at a time, type 2 error decreases. When the third and fifth features are added, it can be seen that the type 1 error increases. But with the highly correlated 9 features, which are the best combination of features, we can see a decrease in type 1 and type 2 error. This is the tipping point, where further adding features that do not clearly distinguish between organic and other segment, gives an increase in type 1 error. With regard to the question on when to stop adding features in order to enrich the segment, based on adding features with the AH data set, choosing the best combination of highly correlated features to the organic segment gives the best results based on analysis of type 1 and type 2 error. It can also be seen that after the nine highly correlated features, as additional features are added, there is a high type 2 error and the enrichment becomes less relevant.

Chapter 7

Conclusion and Recommendations

Customer segmentation at AH is done based on business rules. Based on business rules, a fixed number of features $\{X_1....X_M\}$ are used in order to segment the customers. But, it can be seen that there are customers that are not interested in organic products as part of the organic segment based on business rules. Applying the threshold shifting approach, as in Section 4.3, on the first feature as used in the business rules, leads to adding customers that are not interested in organic products. Showing ads to customers based on their interest, increases the likelihood of the customers to click on the ad and buy the product. Also, based on the number of clicks, supplier of the product pays AH. So, showing ads that a customer is interested in increases the probability of profit for AH. The goal is to enrich the existing customer segments at AH.

In order to enrich existing customers segments at AH, adding additional features $\{X_{M+1}....X_D\}$ is proposed. Mathematical formulation is developed to represent customers as observations in the sample space. The observations are realizations of independent and identically distributed random variables. Based, on the mathematical formulation and assumptions, the Gaussian mixture model via the EM algorithm is set up. Initialization of the parameters of the EM algorithm based on business rules is proposed.

Numerical experiments are carried out to observe the enrichment of segments when additional features are added through the synthetic data set and to show that adding additional features provides meaningful enrichment. Features are added one at a time and the effect of adding features is observed based on type 1 and type 2 error. Numerical experiments proved to show that adding additional features help to decrease type 1 and type 2 error and enrich the segment. Based on Case 3 in the numerical experiments, it was also found that the EM algorithm failed to provide meaningful enrichment when the data points were entirely overlapping. Also, different initializations of the EM algorithm was carried out. When type 1 and type 2 error are compared for different initializations, It can be observed that in order to obtain meaningful enrichment, initialization based on business rules gives the best results.

With the AH data set, it was found that adding highly correlated features to the organic segment helps to enrich the segment based on analysing the type 1 and type 2 error. In comparison to the synthetic data set, where features with the same correlation between features in both the distributions are added, proved to provide meaningful enrichment, this is not the case with the AH data set. It is also observed that features with low correlation to the organic segment are added, the type 1 error increases. As additional features are added and we go in higher dimension, it can be seen that type 1 error increases and type 2 error decreases. As mentioned in Section 3.2, the curses of dimensionality kicks in and clustering provides meaningless results. The thesis has helped to show that adding additional features help to enrich existing customer segments. When, the threshold shifting and the EM algorithm approaches are compared, it is seen that the EM algorithm yields better results as type 1 error decreases in order to enrich the segment. So, the approach of using the EM algorithm by adding additional features is already an improvement to the existing business rules approach in order to enrich the segment. The thesis has also helped to answer the following research questions.

The first research question,

How does adding additional features help to enrich existing segments,

can be answered from the numerical experiments carried out on synthetic data set and from AH data set. Experiments on the synthetic and AH data set shows that adding additional features that have a strong correlation to the organic segment, helps to decrease type 1 and type 2 error. From the AH data set, it can be seen that adding the best combination of top correlated features helps in enriching the segment, based on analysing the type 1 and type 2 error.

The subquestion,

What would a good machine learning model be,

was answered from a combination of mathematical formulation, problem description and literature review. Factors such as distance metrics, overlapping of clusters and shape of clusters needed to be taken into account. The EM algorithm takes into account the covariance structure of the clusters and assigns a probability with which each data point can belong to either of the clusters. Further, the EM algorithm provided good results with the synthetic data set for enrichment of the customers based on comparing type 1 and type 2 error. Also, with the AH data set, when top nine correlated features are added, the algorithm is able to decrease type 1 and type 2 error.

The next subquestion,

What are the relevant features in order to segment customers,

is not a question that can be clearly answered. With the synthetic and the AH data set, it can be seen that adding highly correlated features or features that distinguish between organic and other segment, help to enrich the segment. Based on the synthetic data set, features that have same covariance to organic and other segment, still help to enrich the segment. Furthermore, based on the analysis with the synthetic data set, results on when not to add a particular feature has not been established. With the AH data set, it can be observed that adding highly correlated features to the organic segment help to decrease type 1 and type 2 error.

The next subquestion,

what is the usability of the machine learning model,

can be answered from the approach taken to initialize the parameters. The EM algorithm requires input parameters and initialization of the parameters is done based on business rules. Therefore, the usability of the model from a business perspective is straightforward. The output of the algorithm is either a customer is in the organic segment or not, based on the density estimation of belonging to both the clusters with a probability.

The next question to be addressed is *How does type 1 and type 2 error affect and impact the enrichment of the model,*

can be answered based on the synthetic and AH data set. Experiments on the synthetic and AH data set shows that adding additional features that have a strong correlation to the organic segment, helps to decrease type 1 and type 2 error and enrich the segment. With the AH data set, as one goes in higher dimension, the type 1 error increases and type 2 error decreases. So, there is a very high enrichment at the cost of more customers that are not interested in organic products as part of the segment.

7.1 Reflection

There are several points in this thesis to reflect on, based on which recommendation for future research and for the business can be given. One of the main assumptions that was considered was that the data is normally distributed. With the AH data set the distribution of the data set is not known. Also, the underlying ground truth, that is customers labelled as organic or other, is also not known. In order to perform an offline evaluation of the AH data set, the customers are labelled. This is done by clustering the customers in a high dimensional feature space with thirty one features. Based on the analysis of the clusters, it is assumed that clusters with high mean of organic products are the organic customers. So, conclusions from the AH data set are drawn assuming the ground truth labels obtained from clustering.

7.2 Recommendation for Business

With the AH data set, due to the lack of availability of underlying ground truth, an underlying ground truth was established in order to evaluate the model. To obtain conclusive results, that is if the customers that are added to existing segments are actually interested in organic products, A/B testing the results on customers is recommended. A/B testing is showing two versions of the same web page to determine which one performs better. For instance, 50% of visitors are shown version A ("Control") and the other 50% of visitors are shown version B ("Variant"). The number of clicks on the product are compared in both the versions.

The current use case used with the AH data set is organic products, this approach can be extended to other use cases. A framework, based on AH and synthetic data set, in order to enrich customers is developed. Within this framework, features based on business rules, additional features derived from the AH data set and features based on business knowledge are selected. Selection of features from the AH data set shows that, adding highly correlated features to the organic segment reduces type 1 and type 2 error. The parameters of the EM algorithm are initialised based on business rules. The current framework that was developed gives better enrichment in comparison to the business rules, based on analysis of type 1 and type 2 error. In order to further improve the framework, further research on feature selection and A/B testing is required.

7.3 Recommendation for Further Research

Focus is given on enriching existing customer segments by clustering in the continuous case. The EM algorithm can be extended to cluster data points containing both categorical and continuous variables [57] [58]. Parameters in order to predict, both categorical and continuous variables are defined, based on which the EM algorithm can be applied. In order to further improve and enrich existing customer segments and show personalised ads to customers, a clustering based recommendation system is mentioned in [59]. The paper aims to improve recall and precision based on exploiting preference between users. Combining the approach of the paper with the approach from this thesis, improvement of enrichment of existing customer segments could be possible. With the synthetic data set, it was found that the EM algorithm did not work when the clusters were completely overlapping. In [60], a mixture model based clustering for overlapping clusters is presented. Future work on applying the approach suggested in the paper can be useful in order to provide better clustering results.

Bibliography

[1] J.-J. Jonker, N. Piersma, and D. Van den Poel, "Joint optimization of customer segmentation and marketing policy to maximize long-term profitability," *Expert Systems with Applications*, vol. 27, no. 2, pp. 159–168, 2004.

[2] J. Wasilewski and N. Hurley, "Are you reaching your audience? exploring item exposure over consumer segments in recommender systems," pp. 213–217, 2018.

[3] Y. Sneha and G. Mahadevan, "A study on clustering techniques in recommender systems," in *International Conference on Computational Techniques and Artificial Intelligence*, pp. 97–100, 2011.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[5] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering," in *Proceedings of the fifth international conference on computer and information technology*, vol. 1, pp. 291–324, 2002.

[6] A. Vellido, P. Lisboa, and K. Meehan, "Segmentation of the on-line shopping market using neural networks," *Expert systems with applications*, vol. 17, no. 4, pp. 303–314, 1999.

[7] E. Bair, "Semi-supervised clustering methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 5, pp. 349–361, 2013.

[8] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. d. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PloS one*, vol. 14, no. 1, 2019.

[9] X. Jin, J. Han, C. Sammut, *et al.*, "Partitional clustering.," 2010.

[10] C. P. Ezenkwu, S. Ozuomba, and C. Kalu, "Application of k-means algorithm for efficient customer segmentation: a strategy for targeted customer services," 2015.

[11] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer segmentation using k-means clustering," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 135–139, IEEE, 2018.

[12] I. Smeureanu, G. Ruxanda, and L. M. Badea, "Customer segmentation in private banking sector using machine learning techniques," *Journal of Business Economics and Management*, vol. 14, no. 5, pp. 923–939, 2013.

[13] M. Baranwal and S. M. Salapaka, "Weighted kernel deterministic annealing: A maximum-entropy principle approach for shape clustering," in *2018 Indian Control Conference (ICC)*, pp. 1–6, IEEE, 2018.

[14] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[15] Y. Zhang, T. Bouadi, and A. Martin, "An empirical study to determine the optimal k in ek-nnclus method," in *International Conference on Belief Functions*, pp. 260–268, Springer, 2018.

[16] F. Amer Jid Almahri, D. Bell, and M. Arzoky, “Personas design for conversational systems in education,” 2019.

[17] E. Schubert and P. J. Rousseeuw, “Faster k-medoids clustering: improving the pam, clara, and clarans algorithms,” in *International Conference on Similarity Search and Applications*, pp. 171–187, Springer, 2019.

[18] J. K. Parker, L. O. Hall, and J. C. Bezdek, “Comparison of scalable fuzzy clustering methods,” in *2012 IEEE International Conference on Fuzzy Systems*, pp. 1–9, IEEE, 2012.

[19] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[20] X. Zhao and W. Xu, “An extended affinity propagation clustering method based on different data density types,” *Computational intelligence and neuroscience*, vol. 2015, 2015.

[21] R. Cai, Z. Zhang, A. K. Tung, C. Dai, and Z. Hao, “A general framework of hierarchical clustering and its applications,” *Information Sciences*, vol. 272, pp. 29–48, 2014.

[22] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: An efficient data clustering method for very large databases,” in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’96, (New York, NY, USA), p. 103–114, Association for Computing Machinery, 1996.

[23] S. Guha, R. Rastogi, and K. Shim, “Cure: an efficient clustering algorithm for large databases,” *Information systems*, vol. 26, no. 1, pp. 35–58, 2001.

[24] N. Senthilkumaran and R. Rajesh, “Image segmentation-a survey of soft computing approaches,” in *2009 International Conference on Advances in Recent Technologies in Communication and Computing*, pp. 844–846, IEEE, 2009.

[25] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise.,” in *Kdd*, vol. 96, pp. 226–231, 1996.

[26] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: ordering points to identify the clustering structure,” *ACM Sigmod record*, vol. 28, no. 2, pp. 49–60, 1999.

[27] A. Hinneburg, D. A. Keim, *et al.*, “An efficient approach to clustering in large multimedia databases with noise,” in *KDD*, vol. 98, pp. 58–65, 1998.

[28] S. Zhong and J. Ghosh, “A unified framework for model-based clustering,” *Journal of machine learning research*, vol. 4, no. Nov, pp. 1001–1037, 2003.

[29] V. Melnykov, R. Maitra, *et al.*, “Finite mixture models and model-based clustering,” *Statistics Surveys*, vol. 4, pp. 80–116, 2010.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[31] L. Xu, “Unsupervised learning by em algorithm based on finite mixture of gaussians,” in *World Congress on Neural Networks (Portland, OR)*, vol. 2, pp. 431–434, 1993.

[32] C. J. Wu, “On the convergence properties of the em algorithm,” *The Annals of statistics*, pp. 95–103, 1983.

[33] G. J. McLachlan, T. Krishnan, and S. K. Ng, “The em algorithm,” tech. rep., Papers/Humboldt-Universität Berlin, Center for Applied Statistics and . . . , 2004.

[34] A. Adam and H. Blockeel, “Dealing with overlapping clustering: A constraint-based approach to algorithm selection.,” in *MetaSel@ PKDD/ECML*, pp. 43–54, 2015.

[35] S. Wang, H. Lu, and Z. Liang, “A theoretical solution to map-em partial volume segmentation of medical images,” *International journal of imaging systems and technology*, vol. 19, no. 2, pp. 111–119, 2009.

[36] M. J. Lindstrom and D. M. Bates, “Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data,” *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1014–1022, 1988.

[37] S. M. Diffey, A. B. Smith, A. Welsh, and B. R. Cullis, “A new reml (parameter expanded) em algorithm for linear mixed models,” *Australian & New Zealand Journal of Statistics*, vol. 59, no. 4, pp. 433–448, 2017.

[38] J. Kay, “The em algorithm in medical imaging,” *Statistical methods in medical research*, vol. 6, no. 1, pp. 55–75, 1997.

[39] J. Bongard, “Probabilistic robotics. sebastian thrun, wolfram burgard, and dieter fox.(2005, mit press.) 647 pages,” 2008.

[40] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, “The variational approximation for bayesian inference,” *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[41] D. H. Fisher, “Knowledge acquisition via incremental conceptual clustering,” *Machine learning*, vol. 2, no. 2, pp. 139–172, 1987.

[42] A. Y. Al-Omary and M. S. Jamil, “A new approach of clustering based machine-learning algorithm,” *Knowledge-Based Systems*, vol. 19, no. 4, pp. 248–258, 2006.

[43] S.-O. Map and T. Kohonen, “Self-organizing map,” *Proceedings of the IEEE*, vol. 78, pp. 1464–1480, 1990.

[44] D. Xu and Y. Tian, “A comprehensive survey of clustering algorithms,” *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[45] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[46] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 2015.

[47] L. Parsons, E. Haque, and H. Liu, “Subspace clustering for high dimensional data: a review,” *Acm Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[48] J. Mao and A. K. Jain, “A self-organizing network for hyperellipsoidal clustering (hec),” *Ieee transactions on neural networks*, vol. 7, no. 1, pp. 16–29, 1996.

[49] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[50] R. Xu and D. Wunsch, “H.: Survey of clustering algorithms,” *IEEE transactions on neural networks*, vol. 16, pp. 645–678, 2005.

[51] P. Legendre and L. F. Legendre, *Numerical ecology*. Elsevier, 2012.

[52] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[53] S. Wachter, “Normative challenges of identification in the internet of things: Privacy, profiling, discrimination, and the gdpr,” *Computer law & security review*, vol. 34, no. 3, pp. 436–449, 2018.

[54] C. Couvreur, “The em algorithm: A guided tour,” in *Computer intensive methods in control and signal processing*, pp. 209–222, Springer, 1997.

[55] V. A. Fajardo and J. Liang, “On the em-tau algorithm: a new em-style algorithm with partial e-steps,” *arXiv preprint arXiv:1711.07814*, 2017.

[56] C. Biernacki, G. Celeux, and G. Govaert, “Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models,” *Computational Statistics & Data Analysis*, vol. 41, no. 3-4, pp. 561–575, 2003.

[57] L. Hunt and M. Jorgensen, “Mixture model clustering for mixed data with missing information,” *Computational Statistics & Data Analysis*, vol. 41, no. 3-4, pp. 429–440, 2003.

[58] S. Cang and H. Yu, “A probabilty neural network for continuous and categorical data,” *IFAC Proceedings Volumes*, vol. 38, no. 1, pp. 203–208, 2005.

[59] C. Tran, J.-Y. Kim, W.-Y. Shin, and S.-W. Kim, “Clustering-based collaborative filtering using an incentivized/penalized user model,” *IEEE Access*, vol. 7, pp. 62115–62125, 2019.

[60] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. J. Mooney, “Model-based overlapping clustering,” in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 532–537, 2005.