SYNOPSES ON CLUSTERING A MARKETING DATA OF A BANKING INSTITUTION

Abstract

Unsupervised learning is a part of machine learning, which in turn is a part of Artificial Intelligence itself. This pdf gives a brief example on how a dataset from a marketing campaign of a banking institution would be handled without the creation of unnecessary assumption of unknown values. Data clustering distinguishes by the absence of category information. Basically structure in data is finding in clustering and it has long history in scientific field. The formal study of learning systems is deduced from Machine learning; which is a field of research. It has found to be highly interdisciplinary field which acquires and constructs upon ideas from statistics, computer science (engineering), optimization theory, and numerous other disciplines of science and mathematics.

Unsupervised Learning-

Unsupervised learning is a branch of machine learning that is used to find underlying patterns in data and is often used in exploratory data analysis. Unsupervised learning does not use labelled data like supervised learning, but instead focuses on the data's features. Labelled training data has a corresponding output for each input. When using unsupervised learning, we are not concerned with the targeted outputs because the goal of the algorithm is to find relationships within the data and group data points based on the input data alone. Supervised learning is concerned with labelled data in order to make predictions, but unsupervised learning is not.

The goal of unsupervised learning algorithms is to analyse data and find important features. Unsupervised learning will often find subgroups or hidden patterns within the dataset that a human observer may not pick up on. With a complex dataset, the subgroups may not be so easy to find. This is where unsupervised learning can help us.

Cleaning of dataset (Handling the null values)-

It is necessary to clean the dataset before we train our model.

Method 1- Deleting the record

This method is only used when we have a really big dataset and only a very small percentage of the dataset has missing values. This method maintains the accuracy of the dataset and values like mean, median and mode aren't affected.

Method 2- Creating a new model

This method requires much more time and computational effort as compared to the other methods. Here we take the row with the missing data as the test data and the remaining data as the training data to train our model. We then give the row with the missing values as an input and find those missing values as the output. This is done for all rows with missing data. This way, our model, after training through the rows with complete data, predicts the value of the missing data. This method is generally not used because it is highly time and effort consuming. Hence, it is only used for smaller data sets.

Method 3- Statistical methods

This is one of the best methods used for removing null values from a fairly big dataset. Here we could use the mean, median or mode to replace the null values on the basis of the already present data. Here, we can iterate through our model and find each null value for any particular column and find which cluster does that row fall into and replace the null value by a statistically derived value for that particular cluster.

Clustering Data-

Methods-

Affinity Propagation

Agglomerative Clustering

BIRCH

DBSCAN

K-Means

Mini-Batch K-Means

Mean Shift

OPTICS

Spectral Clustering

Mixture of Gaussians

Affinity Propagation-

Affinity Propagation was first published in 2007 by Brendan Frey and Delbert Dueck in Science. In contrast to other traditional clustering methods, Affinity Propagation does not require you to specify the number of clusters. In layman's terms, in Affinity Propagation, each data point sends messages to all other points informing its targets of each target's relative attractiveness to the sender. Each target then responds to all senders with a reply informing each sender of its availability to associate with the sender, given the attractiveness of the messages that it has received from all other senders. Senders reply to the targets with messages informing each target of the target's revised relative attractiveness to the sender, given the availability messages it has received from all targets. The message-passing procedure proceeds until a consensus is reached. Once the sender is associated with one of its targets, that target becomes the point's exemplar. All points with the same exemplar are placed in the same cluster.

Agglomerative Clustering-

Agglomerative Clustering or bottom-up clustering essentially started from an individual cluster (each data point is considered as an individual cluster, also called leaf), then every cluster calculates their distance with each other. The two clusters with the shortest distance with each other would **merge** creating what we called node. Newly formed clusters once again calculating the member of their cluster distance with another cluster outside of their cluster. The process is repeated

until all the data points assigned to one cluster called root. The result is a tree-based representation of the objects called dendrogram.

BIRCH Clustering-

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset.

BIRCH summarizes large datasets into smaller, dense regions called Clustering Feature (CF) entries. Formally, a Clustering Feature entry is defined as an ordered triple, (N, LS, SS) where 'N' is the number of data points in the cluster, 'LS' is the linear sum of the data points and 'SS' is the squared sum of the data points in the cluster. It is possible for a CF entry to be composed of other CF entries.

DBSCAN Clustering-

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It groups 'densely grouped' data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. The most exciting feature of DBSCAN clustering is that it is robust to outliers. It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

K-Means Clustering-

K-means clustering uses "centroids", K different randomly-initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all of the points assigned to it. Then the process repeats: every point is assigned to its nearest centroid, centroids are moved to the average of points assigned to it. The algorithm is done when no point changes assigned centroid.

The algorithm looks a little bit like-

Initialize K random centroids.

You could pick K random data points and make those your starting points.

Otherwise, you pick K random values for each variable.

For every data point, look at which centroid is nearest to it.

Using some sort of measurement like Euclidean or Cosine distance.

Assign the data point to the nearest centroid.

For every centroid, move the centroid to the average of the points assigned to that centroid.

Repeat the last three steps until the centroid assignment no longer changes.

The algorithm is said to have "converged" once there are no more changes.

Mini Batch K-Means Clustering-

Mini-batch k-means works similarly to the k-means algorithm. The main difference is that in mini-batch k-means the most computationally costly step is conducted on only a random sample of observations as opposed to all observations. This approach can significantly reduce the time required for the algorithm to find convergence (i.e. fit the data) with only a small cost in quality.

Mean Shift Clustering-

Mean shift is falling under the category of a clustering algorithm in contrast of Unsupervised learning that assigns the data points to the clusters iteratively by shifting points towards the mode (mode is the highest density of data points in the region, in the context of the Mean shift). As such, it is also known as the Mode-seeking algorithm. Mean-shift algorithm has applications in the field of image processing and computer vision.

OPTICS Clustering-

OPTICS Clustering stands for Ordering Points To Identify Cluster Structure. It draws inspiration from the DBSCAN clustering algorithm. It adds two more terms to the concepts of DBSCAN clustering. They are:-

1) Core Distance: It is the minimum value of radius required to classify a given point as a core point If the given point is not a Core point, then it's Core Distance is undefined.

2) Reachability Distance: It is defined with respect to another data point q(Let). The Reachability distance between a point p and q is the maximum of the Core Distance of p and the Euclidean Distance (or some other distance metric) between p and q. Note that The Reachability Distance is not defined if q is not a Core point.

This clustering technique is different from other clustering techniques in the sense that this technique does not explicitly segment the data into clusters. Instead, it produces a visualization of Reachability distances and uses this visualization to cluster the data.

Spectral Clustering-

Spectral clustering is an EDA technique that reduces complex multidimensional datasets into clusters of similar data in rarer dimensions. The main outline is to cluster the all spectrum of unorganized data points into multiple groups based upon their uniqueness "Spectral clustering is one of the most popular forms of multivariate statistical analysis" 'Spectral Clustering uses the connectivity approach to clustering', wherein communities of nodes (i.e. data points) that are connected or immediately next to each other are identified in a graph. The nodes are then mapped to a low-dimensional space that can be easily segregated to form clusters.

Gaussian Mixture Model-

Gaussian Mixture Models (GMMs) assume that there are a certain number of Gaussian distributions, and each of these distributions represent a cluster. Hence, a Gaussian Mixture Model tends to group the data points belonging to a single distribution together.

Cleaning choice-

For the given data, I would use the method of **deleting the record** as inaccurate values in the banking sector are not acceptable. Hence, however small the dataset might be, we just cannot use any other method.

As for the deletion method, we delete all the rows (entries) with any row of null value. This clears the dataset quickly and efficiently without any scope of errors that may creep in due to other methods. Since, it is a banking record, we may assume that it consists of an amazingly high amount of entries and won't be affected much if some of it gets erased.

Clustering choice-

Since, the given data is a bank record, it can be considered to be a humungous dataset. Clustering this dataset using Agglomerative or K-means method for clustering would not be very efficient. According to me, the method that should be used is the **BIRCH clustering** method.

BIRCH- Balanced Iterative Reducing and Clustering using Hierarchies. This method is used when we have an exceptionally big dataset. Here, initially, clustering is done exactly like the K-means method where the cluster is divided into 'K' number of clusters, instead, here the best value of 'K' is also decided by the computer itself. The computer itself finds the 'Elbow' in the elbow plot of variation of the dataset for each value of 'K'. This way the value of 'K' gets determined and the data gets divided into tight clusters. These clusters are then clustered using the hierarchical method of clustering. For hierarchical clustering, we use the centroid of these tight clusters as the data points and start clustering the closest of the clusters. We then replace each pair with a single cluster and such pairing is done till only one cluster remains.

How to determine the number of clusters-

In BIRCH clustering, we actually do not need to calculate and input the number of clusters like we would have to in K-means clustering. Yet, to describe the method how the computer itself decides the 'K' value, it is done on a trial and error basis. 'K' values starting from '1' are taken and a graph of reduction in sum of variation versus 'K' value is made (Elbow graph). With the increase in the number of clusters, the sum of variation always decreases. The value of 'K' for which there is a sharp decrease in the summation of variation is the most optimum value.

Conclusion-

To conclude, the main goal of unsupervised learning is to discover hidden and interesting patterns in unlabelled data. Unlike supervised learning, unsupervised learning methods cannot be directly applied to a regression or a classification problem as one has no idea what the values for the output might be. It is best used if you want to find patterns but don't know exactly what you're looking for.