Customer Segmentation

Josep Curto Díaz, Adjunct Professor[a]

[a]IE Business School, Madrid, 28006, Spain, jcurto@faculty.ie.edu

This version was compiled on April 15, 2019

This technical note introduces what is Customer Segmentation and its benefits and limitations.

customer segmentation | clustering | customer analytics | r

The problem.

- **Problem**: we don't know if we have different types of customers and how to approach them
- **Goals**:
  - We want to understand better our customers
  - We want to have clear criteria to segment our customers
- **Why?** To perform specific actions to improve the customer experience

But,... we have many attributes! How we can choose the relevant attributes? How do they combine to explain our customers?

We will use what is called **customer segmentation**.

Definition. We need a formal definition:

> **Customer segmentation** is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits among many others.

Types. The most common forms of customer segmentation are:

- **Geographic segmentation**: considered as the first step to international marketing, followed by demographic and psychographic segmentation.
- **Demographic segmentation**: based on variables such as age, sex, generation, religion, occupation and/or education level.
- **Firmographic**: based on features such as company size (either in terms of revenue or number of employees), industry sector and/or location (city, country and/or region).
- **Behavioral segmentation**: based on knowledge of, attitude towards, usage rate, response, loyalty status, and/or readiness stage to a product.
- **Psychographic segmentation**: based on the study of activities, interests, and/or opinions (AIOs) of customers.
- **Occasional segmentation**: based on the analysis of occasions (for instance, being thirsty).
- **Segmentation by benefits**: based on RFM, CLV, etc.
- **Cultural segmentation**: based on cultural origin.
- **Multi-variable segmentation**: based on the combination of several techniques and/or attributes.

Comparing customers. We want to know if two customers are similar. To compare customers, we will use their attributes. These attributes are a vector of values (numeric or categorical). Basically, we will translate our purpose (comparing customers) into measure the similarity or dissimilarity between objects (vectors of curstomers attributes) and we will use a distance measure such as Euclidean, Manhattan or Minkowski A distance function returns a lower value for pairs of objects that are more similar to one another.

Euclidean Distance.

$$d_2(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Squared Euclidean Distance.

$$d_2^2(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$

Manhattan Distance.

$$d_1(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

Maximum Distance.

$$d_\infty(x,y) = max_i|x_i - y_i|$$

Minkwoski Distance.

$$d_q(x,y) = \sqrt[q]{\sum_{i=1}^{n}(|x_i - y_i|)^q}$$

Jaccard distance.

$$d(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Canberra distance.

$$d(p,q) = \sum_{i=1}^{n}\frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Levenshtein distance.

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) \text{ if } \min(i,j) = 0, \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Note: there are many distance functions (for example, gower, a general coefficient of Similarity).

Techniques. There are many customer segmentation techniques (all of them are belong to unsupervised learning). This list summarizes the available options:

- Hierarchical clustering: Agglomerative, Divisive
- Partitioning clustering: Partitioning Relocation Clustering, Density-Based Partitioning, Subspace Clustering, Grid-Based Methods
- Other: Co-Occurrence of Categorical, Data Methods, Constraint-Based Clustering, Graph Partitioning, Co-Clustering Techniques

For each type, there are many algorithms:

- Agglomerative: Single Link, Complete Link, Average Link, Ward, BRICH, CURE, ROCK, AGNES
- Divisive: Diana, Mona
- Partitioning Relocation Clustering: EM, SNOB, PAM, CLARA, MCLUST, CLARANS, K-MEANS, AUTOCLASS, FUZZY C-MEANS
- Density-Based Partitioning: DBSCAN, SNN, DENCLUE, DB-CLASO, OPTICS
- Subspace Clustering: ENCLUS, ORCLUS, PROCLUS, OPT-GRID, MAFIA
- Grid-Based Methods: BANG, STING, WAVECLUST, CLIQUE, MAFIA

We just introduce a couple of them. R supports all types of clustering (hierarchical, partitioning,...). More information here: https://cran.r-project.org/web/views/Cluster.html

We will reivew about several aspects in this document:

- Clustering tendency: How to statistically evaluate clustering tendency, i.e. if we can really find clusters in a data set.
- Dimensionality reduction: How to reduce the number of attributes in the clustering, ie. do we need all the attributes? In particular using PCA (Principal Component Analysis)
- Clustering techniques: (1) What is k-means and (2) What is hiearchical clustering
- Clustering validation: How to validate a cluster, i.e. how good is our cluster?

Clustering tendency. **Clustering tendency assessment** determines whether a given dataset contains meaningful clusters (i.e., **non-random structure**).

Hopkins statistic is used to assess the **clustering tendency** of a dataset by measuring the probability that a given dataset is generated by a uniform data distribution. In other words it tests the **spatial randomness** of the data.

Let D be a real dataset. The Hopkins statistic can be calculated as follow:

- Sample uniformly n points $(p_1, \ldots, p_n)$ from D.
- For each point $p_i \in D$, find it's nearest neighbor p_j; then compute the distance between p_i and p_j and denote it as $x_i = dist(p_i, p_j)$
- Generate a simulated dataset ($randomD$) drawn from a random uniform distribution with n points $(q_1, \ldots, q_n)$ and the same variation as the original real dataset D.
- For each point $q_i \in randomD$, find it's nearest neighbor q_j in D; then compute the distance between q_i and q_j and denote it $y_i = dist(q_i, q_j)$
- Calculate the Hopkins statistic (H) as the mean nearest neighbor distance in the random dataset divided by the sum of the mean nearest neighbor distances in the real and across the simulated dataset.

The formula is defined as follow:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

A value of H about 0.5 means that $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i$ are close to each other, and thus the data D is uniformly distributed.

The null and the alternative hypotheses are defined as follow:

- **Null hypothesis**: the dataset D is uniformly distributed (i.e., no meaningful clusters).
- **Alternative hypothesis**: the dataset D is not uniformly distributed (i.e., contains meaningful clusters).

If the value of Hopkins statistic is close to zero, then we can reject the null hypothesis and conclude that the dataset D is significantly a clusterable data set.

Principal Component Analysis. **Principal component analysis** (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.

- It uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation).
- The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations.
- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.
- The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

PCA is useful for eliminating dimensions. That means a **dimensionality reduction technique**.

Clustering techniques.

Kmeans. Given a set of observations $(x1, x2, \ldots, xn)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $S = S1, S2, \ldots, Sk$ so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

where μ_i is the mean of points in S_i.

Hiearchical Clustering. Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom (or otherwise). There are two types of hierarchical clustering, Divisive and Agglomerative.

- **Divisive method**: In this method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation.
- **Agglomerative method**: In this method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single cluster left. The related algorithm is shown below.

Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. The following three methods differ in how the distance between each cluster is measured.

- **Single Linkage**: In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.

$$L(R,S) = min(D(x_{ri}, x_{sj}))$$

where R and S are clusters and x_{ri} and x_{sj} are points in these clusters.

- **Complete Linkage**: In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.

$$L(R,S) = max(D(x_{ri}, x_{sj}))$$

where R and S are clusters and x_{ri} and x_{sj} are points in these clusters.

- **Average Linkage**: In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters "r" and "s" to the left is equal to the average length each arrow between connecting the points of one cluster to the other.

$$L(R,S) = \frac{1}{n_r n_s} \sum_{i}^{n_r} \sum_{j}^{n_s} (D(x_{ri}, x_{sj}))$$

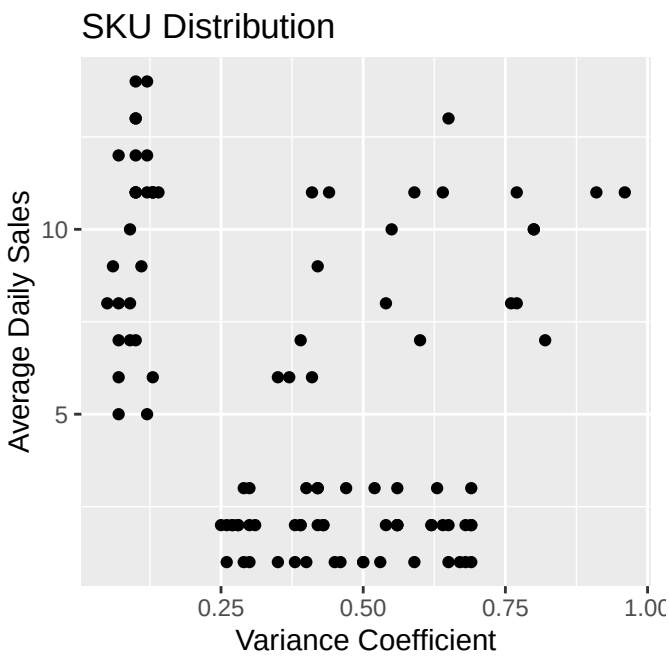
where R and S are clusters and x_{ri} and x_{sj} are points in these clusters.

Clustering validation. The final step is to validate the quality of the cluster. For example, we can use Average Silhouette Analysis to validate whether the clusters has a good structure or not. ¿How to use the value?:

- 0.71 - 1.0: A strong structure has been found.
- 0.51 - 0.70: A reasonable structure has been found.
- 0.26 - 0.50: The structure is weak and could be artificial.
- < 0.25: No substantial structure has been found.

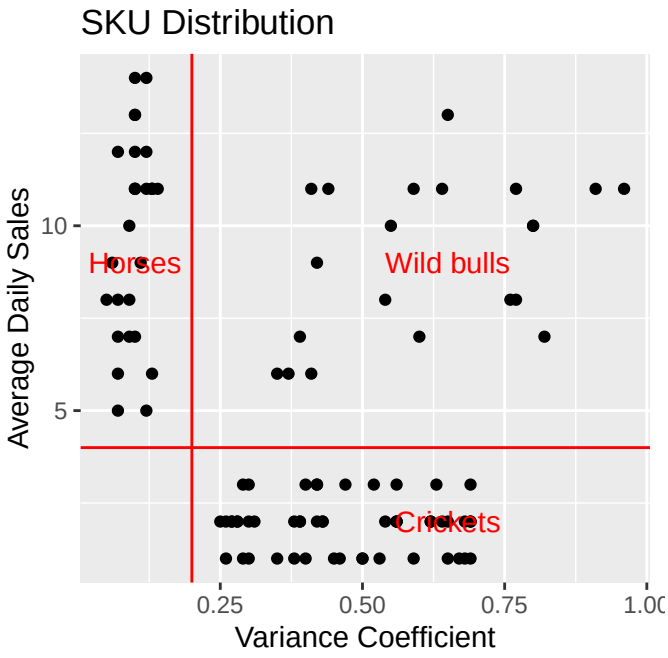
CS is an art. Sometimes data has shape, and shape has meaning,...

```
data <- read.table('data/s8-ex.csv',
                   header = T,sep=',')

library(ggplot2)

ggplot(data, aes(x = CV, y = ADS)) +
  geom_point() +
  labs(x="Variance Coefficient",
       y="Average Daily Sales",
       title="SKU Distribution")
```



We just need to pay attention:

```
ggplot(data, aes(x = CV, y = ADS)) +
  geom_point() +
  labs(x="Variance Coefficient",
       y="Average Daily Sales",
       title="SKU Distribution") +
  geom_hline(yintercept=4, colour="red") +
  geom_vline(xintercept=0.2, colour="red") +
  annotate("text", x =.1, y = 9,
           label = "Horses",colour = "red",
           size=4) +
  annotate("text", x =.65, y = 9,
           label = "Wild bulls",colour = "red",
           size=4) +
  annotate("text", x =.65, y = 2,
           label = "Crickets",colour = "red",
           size=4)
```



Procedure.

- **[BU]** Determine business needs
- **[DU]** Sourcing, Cleaning & Exploration
- **[DP]** Feature Creation (Extract additional information to enrich the set)
- **[DP]** Feature Selection (Reduce to a smaller dataset to speed up computation)
- **[M]** Select Customer Segmentation Technique (test and compare some of them)
- **[M]** Applied Selected Customer Segmentation Technique
- **[E]** Analyze results and adjust parameters
- **[D]** Present and explain the results

Note: Good clustering method requirements are:

- The ability to discover some or all of the hidden clusters.
- Within-cluster similarity and between-cluster dissimilarity.
- Ability to deal with various types of attributes.
- Can deal with noise and outliers.
- Can handle high dimensionality.
- Scalable, Interpretable and usable.

Benefits.

- Customer profiling
- Targeted marketing actions
- Targeted operations

Use Cases.

- Reporting
- Commercial actions: Retention offers, Product promotions, Loyalty rewards
- Operations: Optimise stock levels, store layout
- Pricing: price elasticity
- Strategy: M&A, new products,...

Machine Learning Introduction. K-means is a machine learning algorithm.

Machine Learning refers to a broad set of computer science techniques that let us give computers, as Arthur Samuel put it in 1959, *the ability to learn without being explicitly programmed*. Machine Learning is a subset of Artificial Intelligence (AI). Artificial Intelligence comprises all ML techniques, but it also includes other techniques such as search, symbolic reasoning, logical reasoning, statistical techniques that aren't deep learning based, and behavior-based approaches.

There are several types of AI:

- **Soft, weak, narrow**: AI systems that work in a specific domain. For example, language translation. It is non-sentient artificial intelligence that is focused on one narrow task.
- **Hard/strong/deep**: In the field of artificial intelligence, the most difficult problems are informally known as AI-complete or AI-hard, implying that the difficulty of these computational problems is equivalent to that of solving the central artificial intelligence problem—making computers as intelligent as people, or strong AI.

Machine learning is often split between three main types of learning: **supervised learning**, **unsupervised learning**, and **reinforcement learning**. Knowing the differences between these three types of learning is necessary for any data scientist.

Supervised learning: regroups different techniques which all share the same principles:

- The training dataset contains inputs data (your predictors) and the value you want to predict (which can be numeric or not).
- The model will use the training data to learn a link between the input and the outputs. Underlying idea is that the training data can be generalized and that the model can be used on new data with some accuracy. Some supervised learning algorithms:
- Linear and logistic regression
- Support vector machine
- Naive Bayes
- Neural networks
- Gradient boosting
- Classification trees and random forest

Supervised learning is often used for expert systems in image recognition, speech recognition, forecasting, and in some specific business domain (Targeting, Financial analysis,...).

Unsupervised learning: does not use output data (at least output data that are different from the input). Unsupervised algorithms can be split into different categories:

- Clustering algorithm, such as K-means, hierarchical clustering or mixture models. These algorithms try to discriminate and separate the observations in different groups.
- Dimensionality reduction algorithms (which are mostly unsupervised) such as PCA, ICA or autoencoder. These algorithms find the best representation of the data with fewer dimensions.
- Anomaly detections to find outliers in the data, i.e. observations which do not follow the data set patterns.

Most of the time unsupervised learning algorithms are used to pre-process the data, during the exploratory analysis or to pre-train supervised learning algorithms.

Reinforcement learning: try to find the best ways to earn the greatest reward. Rewards can be winning a game, earning more money or beating other opponents. They present state-of-art results on very human task, for instance, how a computer can beat human in old-school Atari video game.

Reinforcement learnings algorithms follow the different circular steps:

- Given its and the environment's states, the agent will choose the action which will maximize its reward or will explore a new possibility.
- These actions will change the environment's and the agent states.
- They will also be interpreted to give a reward to the agent.
- By performing this loop many times, the agents will improve its behavior.

Reinforcement learning already performs wells on 'small' dynamic system and is definitely to follow for the years to come.

Interesting packages.

- RSKC: An R Package for a Robust and Sparse K-Means Clustering Algorithm
- Clustering Mixed Data Types in R
- Equi-Rank Hierarchical Clustering Validation
- Rdimtools
- CrossClustering: A Partial Clustering Algorithm with Automatic Estimation of the Number of Clusters and Identification of Outliers
- CEC: Cross-Entropy Clustering
- klaR: Classification and Visualization and a introduction
- clustMixType: k-Prototypes Clustering for Mixed Variable-Type Data
- clues: Clustering Method Based on Local
- mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation
- QuClu: Quantile-Based Clustering Algorithms
- clusterlab: Flexible Gaussian Cluster Simulator
- spherical k-Means
- Affinity Propagation clustering
- hierarchical clustering
- hybrid hierarchical clustering
- [Latent Class Analysis (LCA): random LCA
- polytomous variable LCA
- Bayesian LCA
- Mixtools
- PCA
- Topological Data Analysis

References.

- Hwang, H., Jung, T. and Suh, E., 2004. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert systems with applications, 26(2), pp.181-188.
- Kim, S.Y., Jung, T.S., Suh, E.H. and Hwang, H.S., 2006. Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications, 31(1), pp.101-107. -Marcus, C., 1998. A practical yet meaningful approach to customer segmentation. Journal of consumer marketing, 15(5), pp.494-504.
- Chan, C.C.H., 2008. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. Expert systems with applications, 34(4), pp.2754-2762.
- Teichert, T., Shehu, E. and von Wartburg, I., 2008. Customer segmentation revisited: The case of the airline industry. Transportation Research Part A: Policy and Practice, 42(1), pp.227-242.
- Espinoza, M., Joye, C., Belmans, R. and Moor, B.D., 2005. Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series. Power Systems, IEEE Transactions on, 20(3), pp.1622-1630.
- Wu, J. and Lin, Z., 2005, August. Research on customer segmentation model by clustering. In Proceedings of the 7th international conference on Electronic commerce (pp. 316-318). ACM.
- Machauer, A. and Morgner, S., 2001. Segmentation of bank customers by expected benefits and attitudes. International Journal of Bank Marketing, 19(1), pp.6-18.
- Machine Learning with R
- Data Science Live Book
- Peter J. Rousseeuw (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.