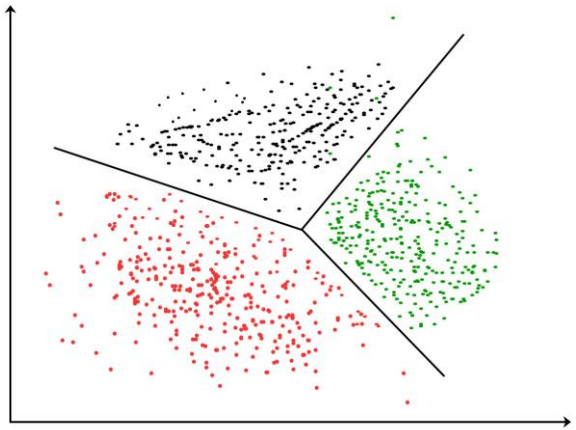
Intro to Clustering

Daro VAN

Clustering for segmentation

Imagine you have a data and you have to apply customer segmentation to this data.

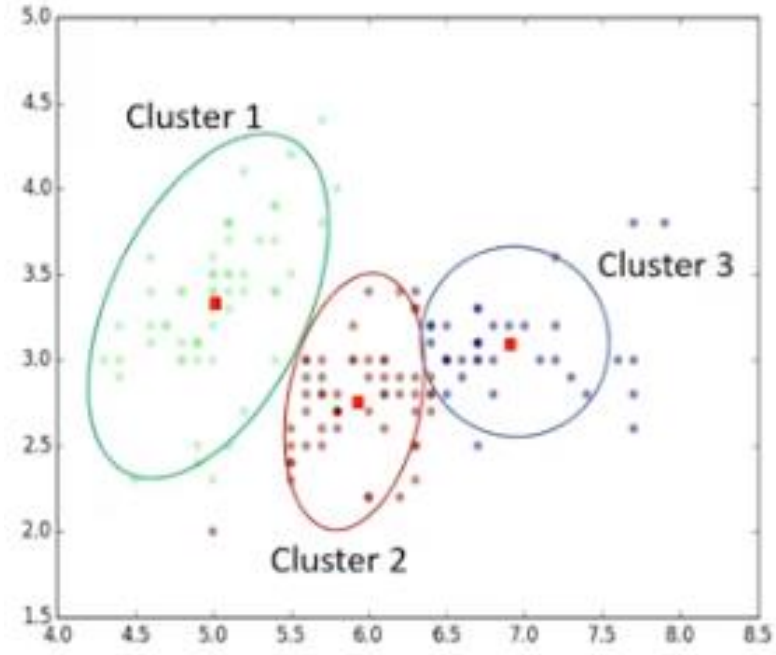


	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	Address	DebtIncomeRatio
0	1	41	2	6	19	0.124	1.073	0.0	NBA001	6.3
1	2	47	1	26	100	4.582	8.218	0.0	NBA021	12.8
2	3	33	2	10	57	6.111	5.802	1.0	NBA013	20.9
3	4	29	2	4	19	0.681	0.516	0.0	NBA009	6.3
4	5	47	1	31	253	9.308	8.908	0.0	NBA008	7.2

Customer segmentation is the practice of partitioning a customer base into groups of individuals that have similar characteristics. It allows the business to target specific groups of customers so as to more effectively allocate marketing resources.

What is clustering

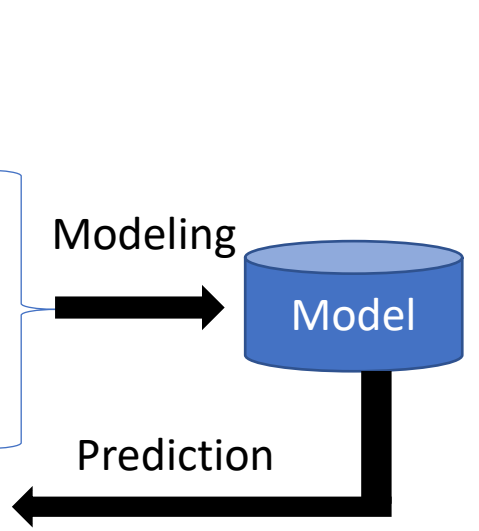
- Clustering means finding the clusters in the data unsupervised.
- Cluster is a group of objects that are similar to other objects in the cluster, and dissimilar to data points in other clusters.



Clustering Vs. Classification

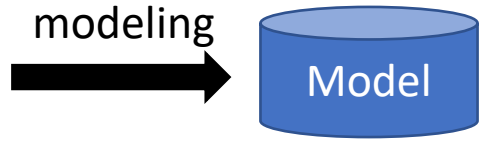
Labeled dataset

	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	DebtIncomeRatio	Clus_km
0	1	41	2	6	19	0.124	1.073	0.0	6.3	1
1	2	47	1	26	100	4.582	8.218	0.0	12.8	2
2	3	33	2	10	57	6.111	5.802	1.0	20.9	1
3	4	29	2	4	19	0.681	0.516	0.0	6.3	1
4	5	47	1	31	253	9.308	8.908	0.0	7.2	



Unlabeled dataset

	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	DebtIncomeRatio	Clus_km
0	1	41	2	6	19	0.124	1.073	0.0	6.3	1
1	2	47	1	26	100	4.582	8.218	0.0	12.8	2
2	3	33	2	10	57	6.111	5.802	1.0	20.9	1
3	4	29	2	4	19	0.681	0.516	0.0	6.3	1
4	5	47	1	31	253	9.308	8.908	0.0	7.2	0



Clustering application

- Retail/marketing
  Identifying buying patterns of customers
  Recommending new books or movies to new customers
- Banking
  Fraud detection in credit card use
  Identifying clusters of customers (e.g. loyal or churn …)
- Insurance
  Fraud detection in claim analysis
  Insurance risk of customers
- Publication
  Auto-categorizing news based on their content
  Recommending similar new article
- Medicine
  Characterizing patient behavior
- And more

What clustering ?

- Exploratory data analysis
- Summary generation
- Outlier detection
- Finding duplicates
- Pre-processing step

Clustering algorithms

- Partitioned-based clustering
  Relatively efficient
  Ex. K-Means, K-Median, Fuzzy c-Means
- Hierarchical clustering
  Produces trees of clusters
  Ex. Agglomerative, Divisive
- Density-Based clustering
  Produces arbitrary shaped clusters
  Ex. DBSCAN