Agenda

Clustering

Association Rules

Text Mining

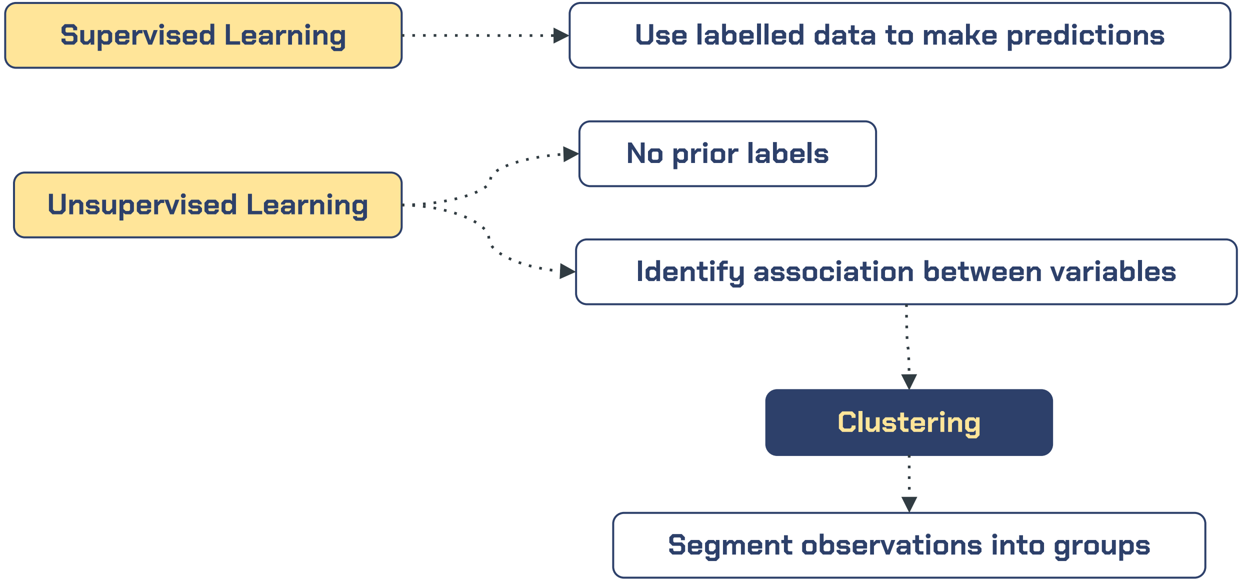
Growth of Data Mining

Explosion of data collection

Advancement in data storage and processing

Affordability and advancement in analytics

Machine Learning methods & Clustering



Supervised Learning

Use labelled data to make predictions

Unsupervised Learning

No prior labels

Identify association between variables

Clustering

Segment observations into groups

Clustering methods

Hierarchical clustering → starts with each observation belonging to its own cluster and then **sequentially merges the most similar clusters** to create a series of nested clusters.

K-means clustering → assigns each observation to one of ***k*** **clusters** in a manner such that the observations assigned to the same cluster are as similar as possible.

Similarity between observations

Measuring similarity

Euclidean distance → **Most common method to measure dissimilarity between observations.**

$$d_{uv} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_q - v_q)^2}$$

$u = [u_1, u_2, u_3 \ldots u_q]$ and $v = [v_1, v_2, v_3 \ldots v_q]$ are 2 sets of observations of a variable.

Matching coefficient →

$$\frac{\text{number of variables with matching value for observations } u \text{ and } v}{\text{total number of variables}}$$

[Encoded categorical variables]

Jaccard's coefficient →

$$\frac{\text{number of variables with matching nonzero value for observations } u \text{ and } v}{(\text{total number of variables}) - (\text{number of variables with matching zero values for observations } u \text{ and } v)}$$

→ Does not consider "0" to be a similarity unlike *Matching coefficient*

Measuring similarity - example

Observation	Female	Married	Loan	Mortgage
1	1	0	0	0
2	0	1	1	1
3	1	1	1	0
4	1	1	0	0
5	1	1	0	0

Observations	1	2	3	4	5
1	1				
2	0	1			
3	0.5	0.5	1		
4	0.75	0.25	0.75	1	
5	0.75	0.25	0.75	1	1

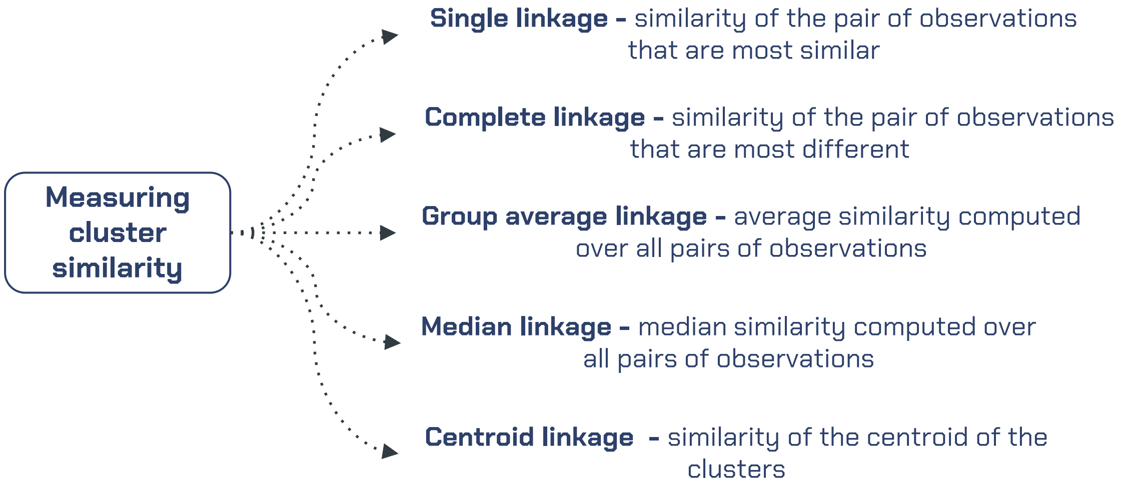
Matching coefficient

Observations	1	2	3	4	5
1	1				
2	0	1			
3	0.33	0.5	1		
4	0.5	0.25	0.67	1	
5	0.5	0.25	0.67	1	1

Jaccard’s coefficient

Hierarchical clustering

Starts with each observation in its own cluster and then iteratively combines the two clusters that are the most similar into a single cluster.



Measuring cluster similarity

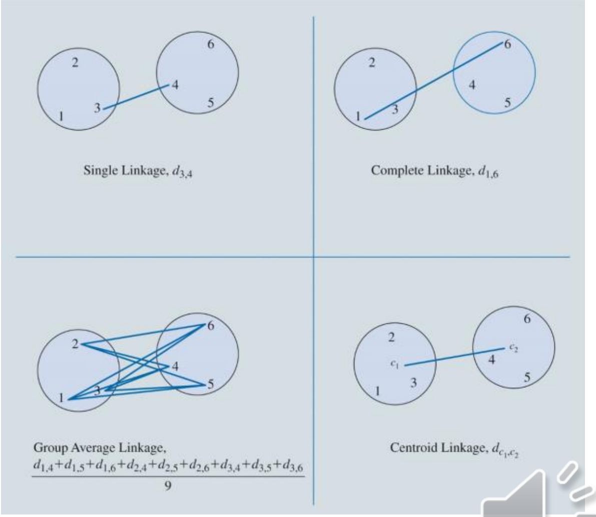
Single linkage - similarity of the pair of observations that are most similar

Complete linkage - similarity of the pair of observations that are most different

Group average linkage - average similarity computed over all pairs of observations

Median linkage - median similarity computed over all pairs of observations

Centroid linkage - similarity of the centroid of the clusters



1
2
3
4
5
6
Single Linkage, $d_{3,4}$
Complete Linkage, $d_{1,6}$
Group Average Linkage, $\frac{d_{1,4}+d_{1,5}+d_{1,6}+d_{2,4}+d_{2,5}+d_{2,6}+d_{3,4}+d_{3,5}+d_{3,6}}{9}$
c_1
c_2
Centroid Linkage, d_{c_1,c_2}

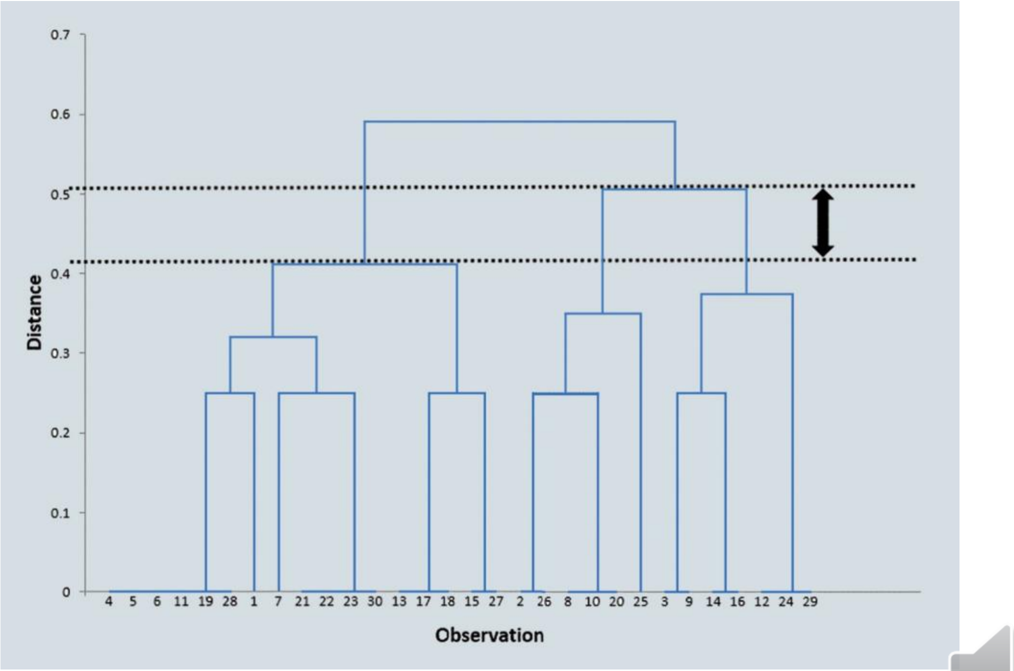
$$d_{1,2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Hierarchical Clustering

Ward's method - merges two clusters such that the dissimilarity of the observations with the resulting single cluster increases as little as possible.

McQuitty's method - considers merging two clusters A and B, the dissimilarity of the resulting cluster AB to any other cluster C is calculated as: ((dissimilarity between A and C) + (dissimilarity between B and C)) divided by 2).

Dendrogram - chart that depicts the set of nested clusters resulting at each step of aggregation

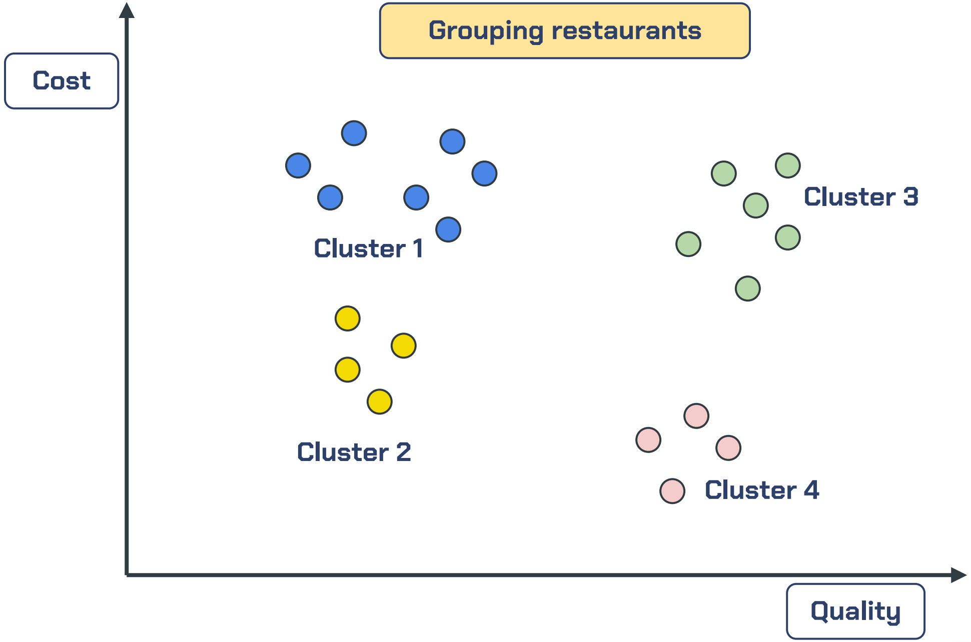


K-means Clustering

Given a value of *k*, the *k*-means algorithm randomly assigns each observation to one of the *k* clusters.

After all observations have been assigned to a cluster, the resulting cluster centroids are calculated.

Using the cluster centroids, all observations are reassigned to the cluster with the closest centroid.



Choosing the Clustering method

Hierarchical Clustering	*k*-Means Clustering
Suitable when we have a **small data set** (e.g., fewer than 500 observations).	Suitable when you know how many clusters you want and you have a **larger data set** (e.g., more than 500 observations).
Convenient method if you want to observe how clusters are **nested**.	Ideal for **quantitative data**

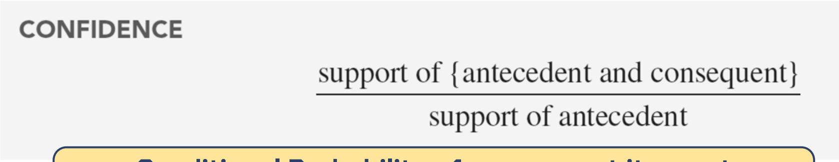
Association rules

If-then statements which convey the **likelihood** of certain variables **occurring together**.

Antecedent: The collection of items (or item set) corresponding to the *if* portion of the rule.

Consequent: The item set corresponding to the *then* portion of the rule.

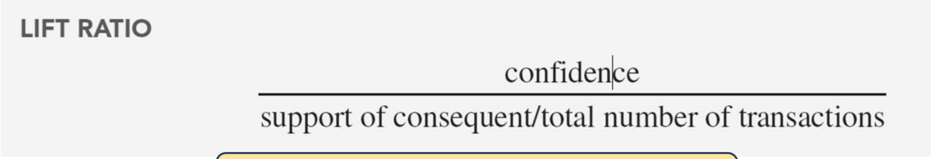
Support count of an item set: Number of transactions in the data that include that item set.



CONFIDENCE

$$\frac{\text{support of \{antecedent and consequent\}}}{\text{support of antecedent}}$$

Conditional Probability of consequent item set occurring given the antecedent



LIFT RATIO

$$\frac{\text{confidence}}{\text{support of consequent/total number of transactions}}$$

Effectiveness the association rule

Example

Transaction	Shopping Cart
1	bread, peanut butter, milk, fruit, jelly
2	bread, jelly, soda, potato chips, milk, fruit, vegetables, peanut butter
3	whipped cream, fruit, chocolate sauce, beer
4	steak, jelly, soda, potato chips, bread, fruit
5	jelly, soda, peanut butter, milk, fruit
6	jelly, soda, potato chips, milk, bread, fruit
7	fruit, soda, potato chips, milk
8	fruit, soda, peanut butter, milk
9	fruit, cheese, yogurt
10	yogurt, vegetables, beer

if **{bread, jelly},** then **{peanut butter}**

Confidence = support { *bread, jelly* and *peanut butter***}/support {** *bread, jelly***}**
= 2/4 = 0.5

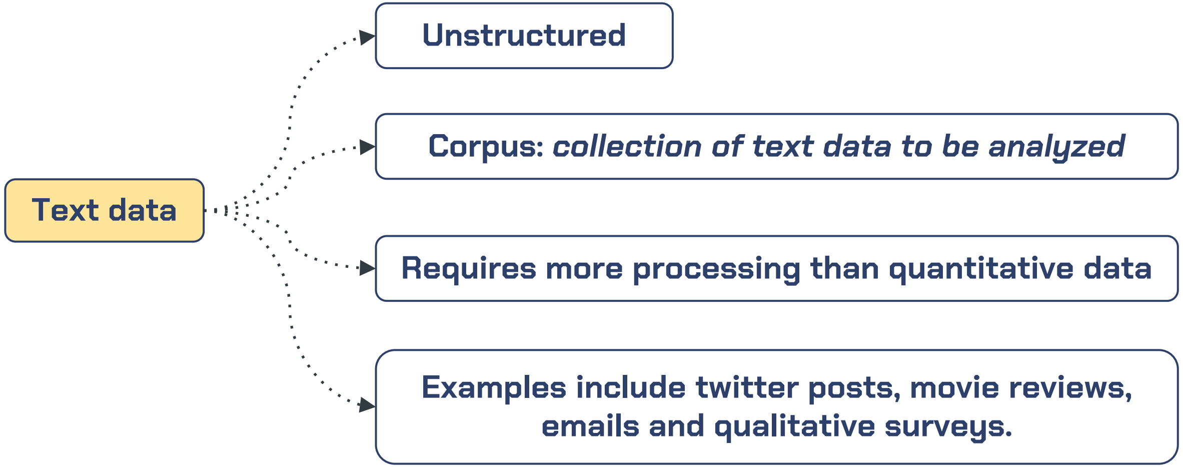
Conditional Probability of consequent item set occurring given the antecedent

Lift ratio = Confidence/(support { *peanut butter***} / total transactions)**
= 0.5/ (4/10) = 1.25

Effectiveness the association rule

Text Mining

Extracting useful information from text



Text data

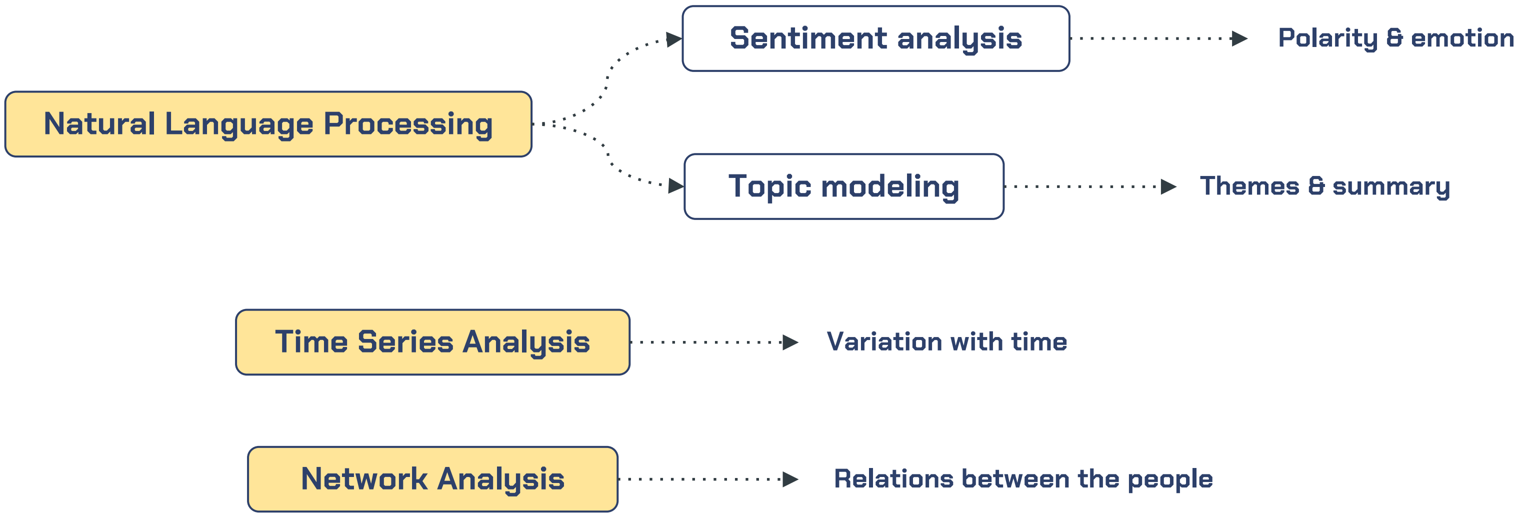
Unstructured

Corpus: *collection of text data to be analyzed*

Requires more processing than quantitative data

Examples include twitter posts, movie reviews, emails and qualitative surveys.

Text mining techniques



Natural Language Processing

Sentiment analysis → Polarity & emotion

Topic modeling → Themes & summary

Time Series Analysis → Variation with time

Network Analysis → Relations between the people

Text mining (NLP) - process

Topic Modeling

- Sentence segmentation
- Word tokenization
- Filter stop words
- Stemming
- Evaluate Term frequency
- Frequency term - document matrix

Sentiment Analysis

- Sentence segmentation
- Word tokenization
- Filter stop words
- Identify “parts of speech”
- Calculate the sentiment

Topic Modeling Example

Concerns
The wi-fi service was horrible. It was slow and cut off several times.
My seat was uncomfortable.
My flight was delayed 2 hours for no apparent reason.
My seat would not recline.
The man at the ticket counter was rude. Service was horrible.
The flight attendant was rude. Service was bad.
My flight was delayed with no explanation.
My drink spilled when the guy in front of me reclined his seat.
My flight was canceled.
The arm rest of my seat was nasty.

	Term						
Document	**Delayed**	**Flight**	**Horrible**	**Recline**	**Rude**	**Seat**	**Service**
1	0	0	1	0	0	0	1
2	0	0	0	0	0	1	0
3	1	1	0	0	0	0	0
4	0	0	0	1	0	1	0
5	0	0	1	0	1	0	1
6	0	1	0	0	1	0	1
7	1	1	0	0	0	0	0
8	0	0	0	1	0	1	0
9	0	1	0	0	0	0	0
10	0	0	0	0	0	1	0

Term-Document matrix

Sentiment Analysis Example

Reviews	Word tokenization and filter	Sentiment
One of the worst movie ever	One, worst, movie, ever	**-0.8**
Boring..	boring	**-0.2**
Great acting weak screenplay	Great, acting, weak, screenplay	**0**
One of the best performance by Actor X	One, best, performance, actor	**1**
Average	average	**0**
One time watch	One, time, watch	**0.1**
Never going to get back the 2 hours I've wasted	Never, going, get, back, 2, hours, wasted	**-0.8**
Wasted potential	Wasted, potential	**-0.9**
This movie makes "Another movie" look like a masterpiece	Movies, makes, another movie, look, masterpiece	**1**
Decent movie	Decent, movie	**0.4**

Summary

Clustering

Association Rules

Text Mining