Descriptive Analytic

Chapter 4 from Analytics in a Big Data World by Bart Baesens

Instructor, Nero Chan Zhen Yu

Forward School

Descriptive Analytics

- In descriptive analytics, the aim is to describe patterns
  - E.g. - customer behavior
- Contrary to predictive analytics, there is no real target variable (e.g., churn or fraud indicator) available
- Hence, descriptive analytics is often referred to as unsupervised learning
  - because there is no target variable to steer the learning process

Forward School

Types of descriptive analytics

- The three most common types of descriptive analytics are

Type of Descriptive Analytics	Explanation	Example
Association rules	Detect frequently occurring patterns between items	Detecting what products are frequently purchased together in a supermarket context Detecting what words frequently co-occur in a text document Detecting what elective courses are frequently chosen together in a university setting
Sequence rules	Detect sequences of events	Detecting sequences of purchase behavior in a supermarket context Detecting sequences of web page visits in a web mining context Detecting sequences of words in a text document
Clustering	Detect homogeneous segments of observations	Differentiate between brands in a marketing portfolio Segment customer population for targeted marketing

Forward School

Association Rule

Forward School

ASSOCIATION RULES

- A Rule-based machine learning method
- Examples of association rules are:
  - If a customer has a car loan and car insurance, then the customer has a checking account in 80% of the cases.
  - If a customer buys spaghetti, then the customer buys red wine in 70 percent of the cases.
  - If a customer visits web page A, then the customer will visit web page B in 90% of the cases.

Forward School

Basic setting of Association Rules

- Association rules typically start from a database of transactions, *D*
- Each transaction consists of a transaction identifier and a set of items (e.g., products, Web pages, courses) $\{i_1, i_2, \ldots, i_n\}$ selected from all - possible items (*I*)
- An association rule is then an implication of the form $X \Rightarrow Y$, whereby $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$
  - *X* is referred to as the rule antecedent, whereas *Y* is referred to as the rule consequent.

Forward School

Example transaction of a supermarket

Transaction Identifier	Items
1	Beer, milk, diapers, baby food
2	Coke, beer, diapers
3	Cigarettes, diapers, baby food
4	Chocolates, diapers, milk, apples
5	Tomatoes, water, apples, beer
6	Spaghetti, diapers, baby food, beer
7	Water, beer, baby food
8	Diapers, baby food, spaghetti
9	Baby food, beer, diapers, milk
10	Apples, wine, baby food

Forward School

Support and confidence

- Two key measures to quantify the strength of an association rule
- The support of an item set is defined as the percentage of total transactions in the database that contains the item set
  - Hence, the rule $X \Rightarrow Y$ has support (s) if $100s\%$ of the transactions in D contain $X \cup Y$. It can be formally defined as follows:

$$support(X \cup Y) = \frac{number\ of\ transactions\ supporting\ (X \cup Y)}{total\ number\ of\ transactions}$$

  - Q: Support for the association rule: Baby food and diapers => beer?
- A frequent item set is one for which the support is higher than a threshold (***minsup***) that is typically specified upfront by the business user or data analyst.
- A lower (higher) support will obviously generate more (less) frequent item sets.

Forward School

Support and confidence

- Two key measures to quantify the strength of an association rule
- The confidence measures the strength of the association and is defined as the conditional probability of the rule consequent, given the rule antecedent.
- The rule $X \Rightarrow Y$ has confidence (c) if 100% of the transactions in D that contain X also contain Y. It can be formally defined as follows:

$$confidence(X \rightarrow Y) = P(Y \mid X) = \frac{support(X \cup Y)}{support(X)}$$

- Again, the data analyst has to specify a minimum confidence (***minconf***) in order for an association rule to be considered interesting

Forward School

Association Rule Mining

- Mining association rules from data is essentially a two-step process as follows:
  - Identification of all item sets having support above ***minsup*** (i.e., “frequent” item sets)
  - Discovery of all derived association rules having confidence above ***minconf***
- Both minsup and minconf must be specified beforehand (by the data analyst)
- The first step is typically performed using the Apriori algorithm

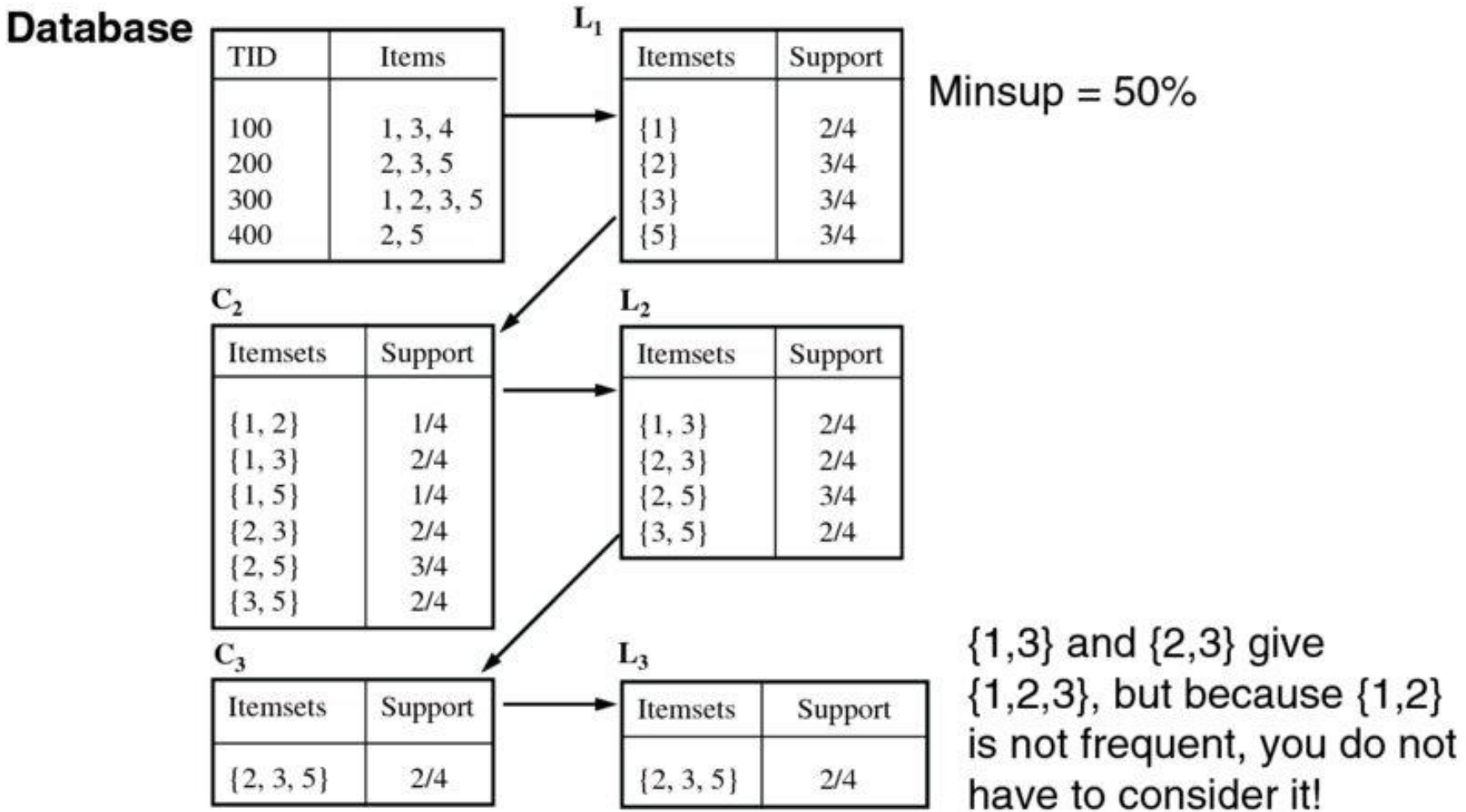
Forward School

Apriori Algorithm

- The basic notion of a priori states that every subset of a frequent item set is frequent as well or, conversely, every superset of an infrequent item set is infrequent.
- This implies that candidate item sets with k items can be found by pairwise joining frequent item sets with k− 1 items and deleting those sets that have infrequent subsets.
- Thanks to this property, the number of candidate subsets to be evaluated can be decreased, which will substantially improve the performance of the algorithm because fewer databases passes will be required
- Once the frequent item sets have been found, the association rules can be generated in a straightforward way, as follows:
  - For each frequent item set k, generate all nonempty subsets of k
  - For every nonempty subset s of k, output the rule s ⇒ k − s if the confidence > minconf
- Note that the confidence can be easily computed using the support values that were obtained during the frequent item set mining.

Forward School

Apriori Algorithm



Result = { {1},{2},{3},{5},{1,3},{2,3},{2,5},{3,5},{2,3,5} }

Forward School

Back to the example: Baby food, diapers and…

- For the frequent item set {baby food, diapers, beer}, the following association rules can be derived:
  - diapers, beer ⇒ baby food [*conf* = 75%]
  - baby food, beer ⇒ diapers [*conf* = 75%]
  - baby food, diapers ⇒ beer [*conf* = 60%]
  - beer ⇒ baby food and diapers [*conf* = 50%]
  - baby food ⇒ diapers and beer [*conf* = 43%]
  - diapers ⇒ baby food and beer [*conf* = 43%]
- If the ***minconf*** is set to 70 percent, only the first two association rules will be kept for further analysis.

Forward School

Lift measure (value)

- Lift measure helps to interpret the importance (i.e. usefulness) of a rule
- While it is a measure for a rule, but you cannot define a minimum lift in the settings similar to minimum support (***minsup***) or minimum confidence (***minconf***)
- The lift is measured as follow:

$$Lift(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X) \times support(Y)}$$

- You can also view it as

$$\text{Lift} = \frac{\text{Rule confidence}}{\text{Prior proportion of the consequent}}$$

- A lift value less (larger) than 1 indicates a negative (positive) dependence or substitution (complementary) effect

Forward School

Example

	Tea	Not Tea	Total
Coffee	150	750	900
Not coffee	50	50	100
Total	200	800	1,000

- Let's now consider the association rule tea ⇒ coffee. The support of this rule is 150/1,000, or 15 percent. The confidence of the rule is 150/200, or 75 percent.
- At first sight, this association rule seems very appealing given its high confidence
  - However, closer inspection reveals that the prior probability of buying coffee equals 900/1,000, or 90 percent
  - Hence, a customer who buys tea is less likely to buy coffee than a customer about whom we have no information
- In our example, the lift value equals 0.89, which clearly indicates the expected substitution effect between coffee and tea

Forward School

Post Processing Association Rules

- Typically, an association rule mining exercise will yield lots of association rules such that post processing will become a key activity. Example steps that can be considered here are:
  - Filter out the trivial rules that contain already known patterns (e.g., buying spaghetti and spaghetti sauce). This should be done in collaboration with a business expert.
  - Perform a sensitivity analysis by varying the ***minsup*** and ***minconf*** values. Especially for rare but profitable items (e.g., Rolex watches), it could be interesting to lower the minsup value and find the interesting associations.
  - Use appropriate visualization facilities (e.g., OLAP based) to find the unexpected rules that might represent novel and actionable behavior in the data.
  - Measure the economic impact (e.g., profit, cost) of the association rules.

Applications of Association Rules

- The most popular application of association rules is market basket analysis.
  - The aim here is to detect which products or services are frequently purchased together by analyzing market baskets
  - Finding these associations can have important implications for targeted marketing (e.g., next best offer), product bundling, store and shelf layout, and/or catalog design.
- Another popular application is recommender systems
  - These are the systems adopted by companies such as Amazon and Netflix to give a recommendation based on past purchases and/or browsing behavior
- What else?

Forward School

Sequence Rules

Forward School

Sequence Rules

- Given a database *D* of customer transactions
  - the problem of mining sequential rules is to find the maximal sequences among all sequences that have certain user-specified minimum support and confidence
- An example could be a sequence of web page visits in a web analytics - setting, as follows:
  - Home page ⇒ Electronics ⇒ Cameras and Camcorders ⇒ Digital Cameras ⇒ Shopping cart ⇒ Order confirmation ⇒ Return to shopping
- It is important to note that a transaction time or sequence field will now be included in the analysis
- Sequence rules are concerned about what items appear at different times (inter-transaction patterns)
  - Whereas association rules are concerned about what items appear together at the same time (intra-transaction patterns)

Forward School

Sequence Rules mining

- To mine the sequence rules, one can again make use of the *a priori* property because if a sequential pattern of length *k* is infrequent, its supersets of length *k* + 1 cannot be frequent.
  Consider the following example of a transactions data set in a web analytics setting (see table on the right). The letters A, B, C, … refer to web pages.
- A sequential version can then be obtained as follows:
  - Session 1: A, B, C
  - Session 2: B, C
  - Session 3: A, C, D
  - Session 4: A, B, D
  - Session 5: D, C, A

Session ID	Page	Sequence
1	A	1
1	B	2
1	C	3
2	B	1
2	C	2
3	A	1
3	C	2
3	D	3
4	A	1
4	B	2
4	D	3
5	D	1
5	C	2
5	A	3

Forward School

Sequence Rule calculation

- One can now calculate the support in two different ways.
- Consider, for example, the sequence rule A ⇒ C.
  - A first approach would be to calculate the support whereby the consequent can appear in any subsequent stage of the sequence. In this case, the support becomes 2/5 (40%).
  - Another approach would be to only consider sessions in which the consequent appears right after the antecedent. In this case, the support becomes 1/5 (20%). A similar reasoning can now be followed for the confidence, which can then be 2/4 (50%) or 1/4 (25%), respectively.

Session ID	Page	Sequence
1	A	1
1	B	2
1	C	3
2	B	1
2	C	2
3	A	1
3	C	2
3	D	3
4	A	1
4	B	2
4	D	3
5	D	1
5	C	2
5	A	3

Forward School

Confidence of a rule

- Remember that the confidence of a rule $A_1 \Rightarrow A_2$
  - is defined as the probability $P(A_2 \mid A_1) = \text{support}(A_1 \cup A_2)/\text{support}(A_1)$.
- For a rule with multiple items, $A_1 \Rightarrow A_2 \Rightarrow \ldots A_{n-1} \Rightarrow A_n$,
  - the confidence is defined as $P(A_n \mid A_1, A_2, \ldots, A_{n-1})$, or
  - $\text{support}(A_1 \cup A_2 \cup \ldots \cup A_{n-1} \cup A_n)/\text{support}(A_1 \cup A_2 \cup \ldots \cup A_{n-1})$.

Forward School

Example for sequential pattern mining

- https://webdocs.cs.ualberta.ca/~zaiane/courses/cau/slides/w4-ex-sol.pdf
- Apply the AprioriAll algorithm to the following customer sequence dataset using minimum support s=33%. Identify the maximal sequence patterns.

S.ID	Sequence
1	<{1 5}{2}{3}{4}>
2	<{1}{3}{4}{3 5}>
3	<{1}{2}{3}{4}>
4	<{1}{3}{5}>
5	<{4}{5}>

Forward School

Segmentation

Forward School

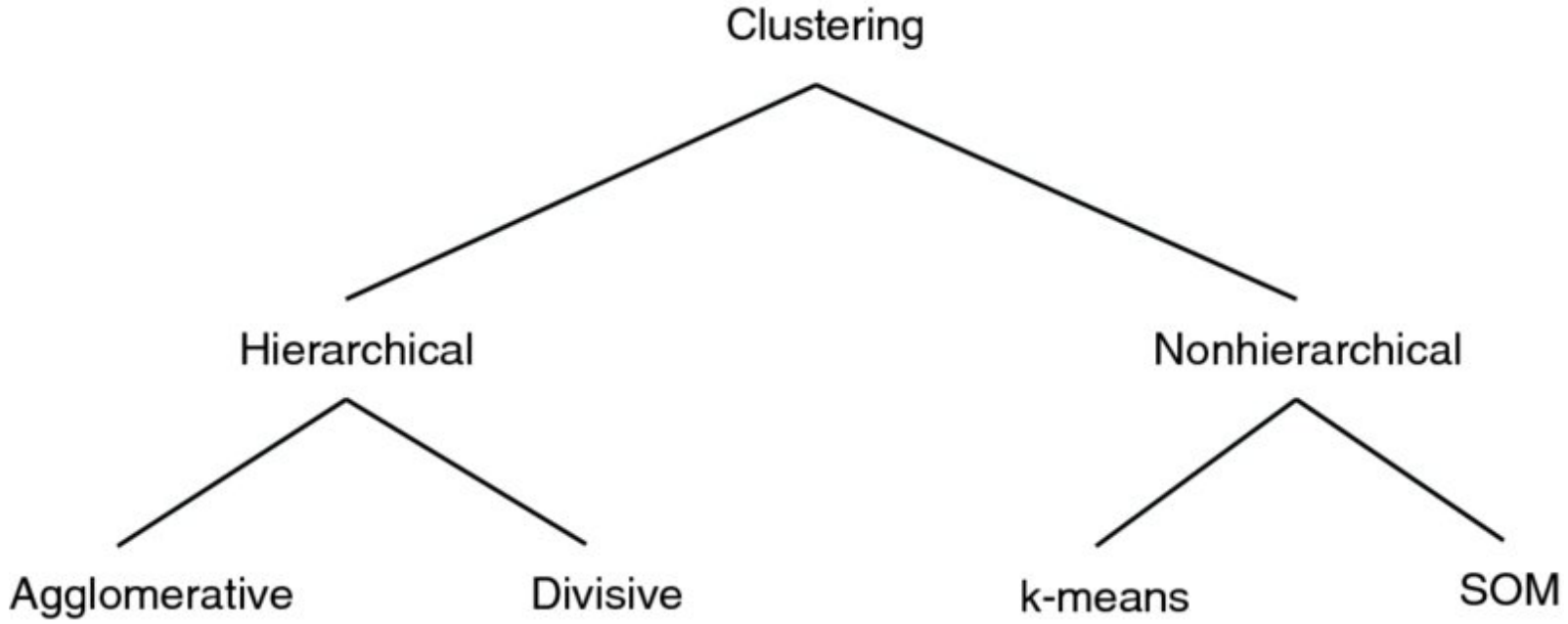
SEGMENTATION

- The aim of segmentation is to split up a set of customer observations into segments such that the homogeneity within a segment is maximized (cohesive) and the heterogeneity between segments is maximized (separated). Popular applications include:
  - Understanding a customer population (e.g., targeted marketing or advertising [mass customization])
  - Efficiently allocating marketing resources
  - Differentiating between brands in a portfolio
  - Identifying the most profitable customers
  - Identifying shopping patterns
  - Identifying the need for new products
- Various types of clustering data can be used, such as demographic, lifestyle, attitudinal, behavioral, RFM, acquisitional, social network, and so on.

Forward School

Clustering

- Clustering techniques can be categorized as either hierarchical or nonhierarchical



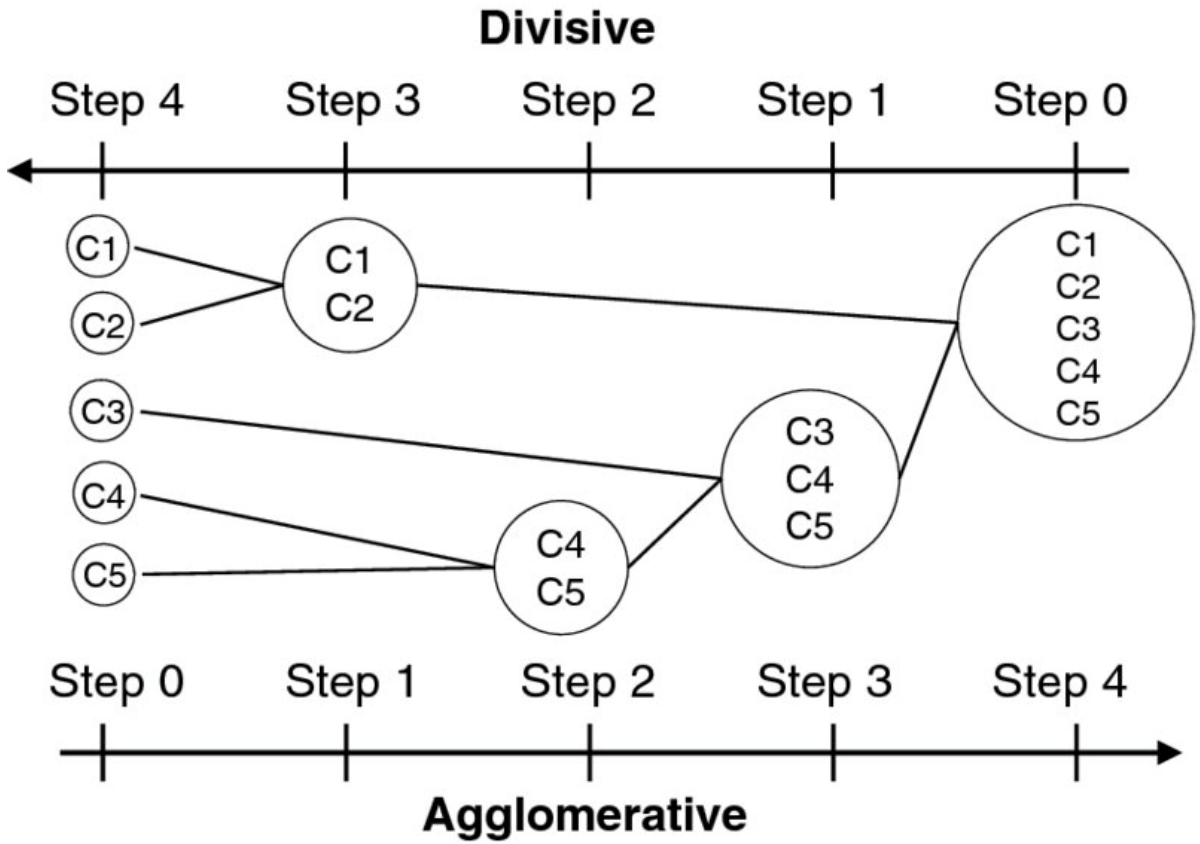
Forward School

HIERARCHICAL CLUSTERING

- Divisive hierarchical clustering
  - starts from the whole data set in one cluster
  - and then breaks this up in each time smaller clusters until one observation per cluster remains
- Agglomerative clustering works the other way around
  - starting from all observations in one cluster
  - and continuing to merge the ones that are most similar until all observations make up one big cluster

Forward School

HIERARCHICAL CLUSTERING



Forward School

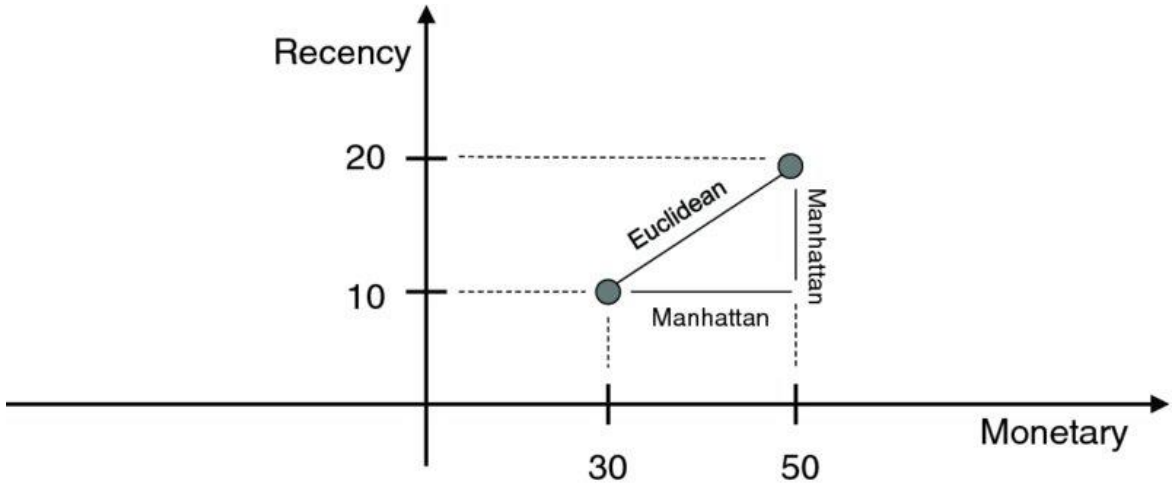
“Rules” to help splitting

- In order to decide on the merger or splitting, a similarity rule is needed.
- Examples of popular similarity rules are the Euclidean distance and Manhattan (city block) distance.
- As an example, both are calculated as follows:

$$\text{Euclidean} : \sqrt{(50-30)^2+(20-10)^2} = 22$$

$$\text{Manhattan} : |50-30| + |20-10| = 30$$

- It is obvious that the Euclidean distance will always be shorter than the Manhattan distance.



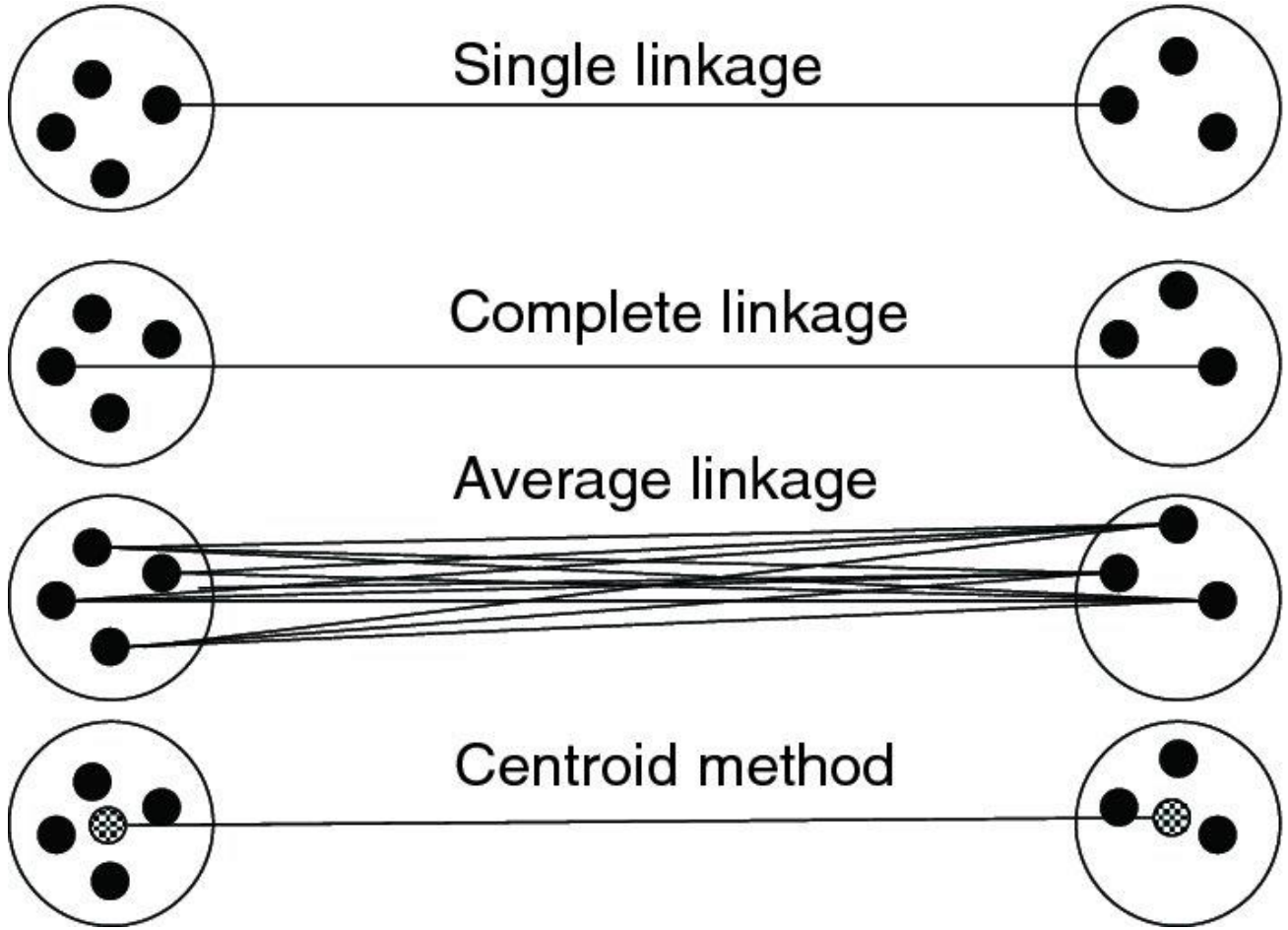
Forward School

Distance calculations

- Various schemes can now be adopted to calculate the distance between two clusters
  - The single linkage method defines the distance between two clusters as the shortest possible distance, or the distance between the two most similar objects.
  - The complete linkage method defines the distance between two clusters as the biggest distance, or the distance between the two most dissimilar objects.
  - The average linkage method calculates the average of all possible distances.
  - The centroid method calculates the distance between the centroids of both clusters.
- Finally, Ward's method merges the pair of clusters that leads to the minimum increase in total within-cluster variance after merging.

Forward School

Schemes for distance calculation



Forward School

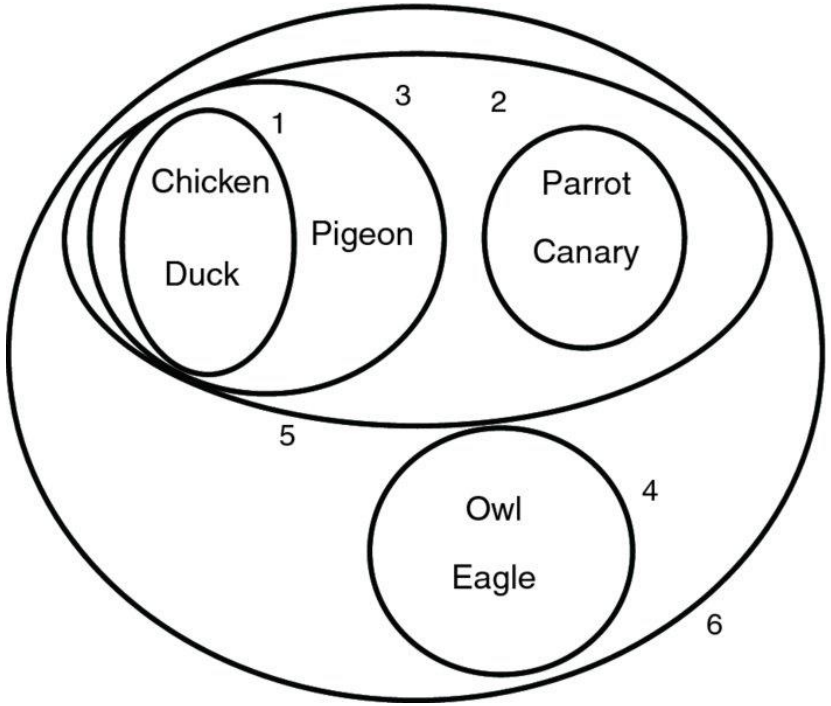
Optimal number of cluster

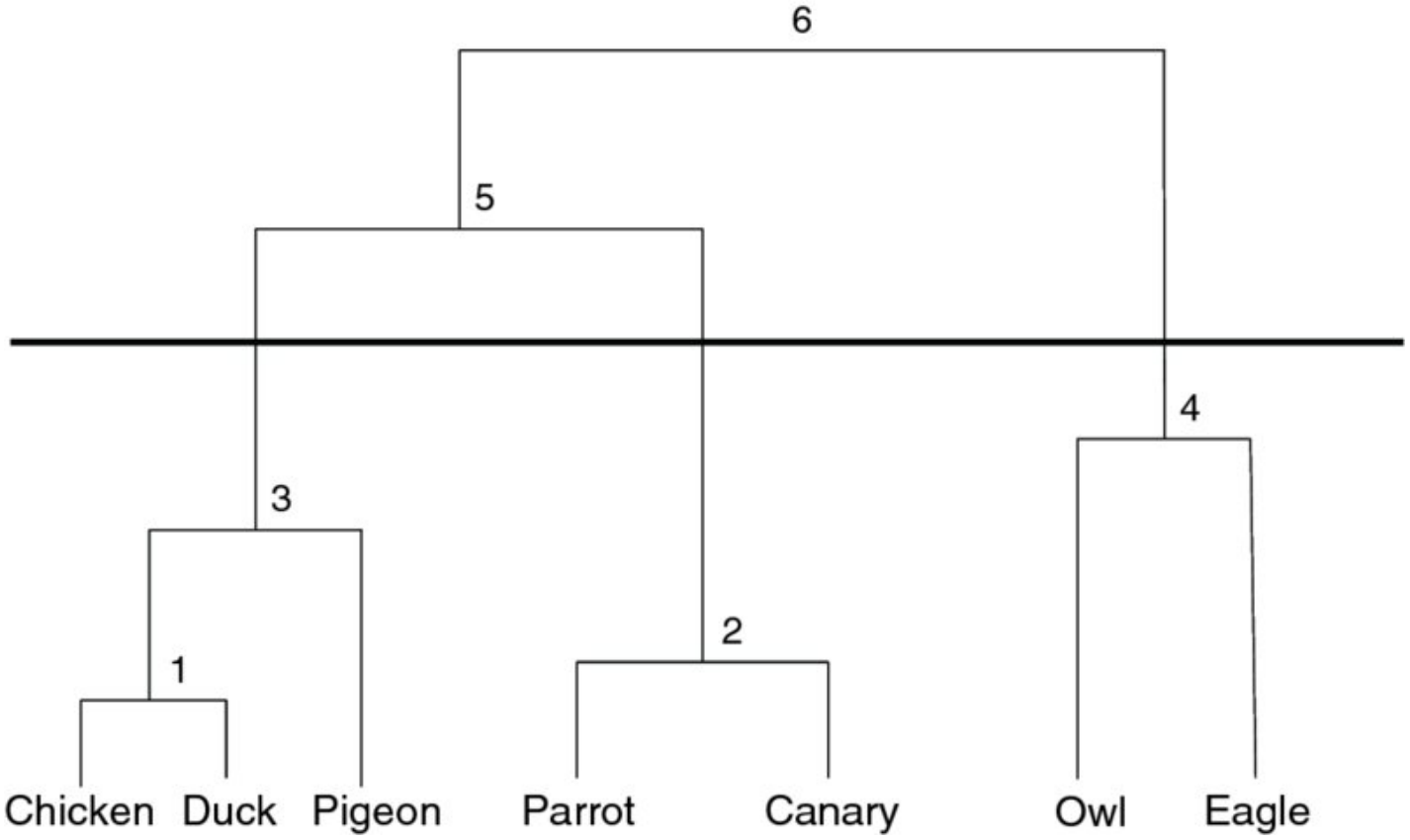
- In order to decide on the optimal number of clusters, one could use a dendrogram or scree plot.
- A dendrogram is a tree-like diagram that records the sequences of merges.
  - The vertical (or horizontal scale) then gives the distance between two clusters amalgamated.
  - One can then cut the dendrogram at the desired level to find the optimal clustering.
- A scree plot is a plot of the distance at which clusters are merged. The elbow point then indicates the optimal clustering.

Forward School

Dendrogram

- **The** numbers indicate the clustering steps.





Forward School

K-means clustering

- *K*-means clustering is a nonhierarchical procedure that works along the following steps:
  - Select *k* observations as initial cluster centroids (seeds).
  - Assign each observation to the cluster that has the closest centroid (for example, in Euclidean sense).
  - When all observations have been assigned, recalculate the positions of the *k* centroids.
  - Repeat until the cluster centroids no longer change.
- A key requirement here is that the number of clusters, *k,* needs to be specified before the start of the analysis. It is also advised to try out different seeds to verify the stability of the clustering solution.

Forward School