Data Warehouse and Data Mining

Dhruv Gupta

dhruv.gupta@ntnu.no

02-February-2021

NTNU | Norwegian University of Science and Technology

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Association Rules
   - Maximal and Closed Itemsets
   - Alternative Methods for Itemset Generation
     - FP-Growth Algorithm
     - ECLAT Algorithm
   - Interestingness Measures

3. Clustering
   - Introduction
   - K-Means Clustering
   - Hierarchical Clustering

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

Administrative

1. Second Assignment
   - Available and due by 19.March.2021.
2. Volunteers for feedback regarding course
   - Interested? Please contact me by email!

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

References for "Association Rules" and "Clustering"

1. Book: Tan et al. *"Introduction to Data Mining"*, 1st Edition, 2006, Pearson Education Inc.
2. Text and images for majority of slides in "Association Rules" subsection are based on the book by Tan et al.

References for "FP-Growth Algorithm"

1. Book: Zaki and Meira. *"Data Mining and Machine Learning: Fundamental Concepts and Algorithms"*, 2nd Edition, 2020, Cambridge University Press.
2. Text and images for some slides in "FP-Growth Algorithm" subsection are based on the book by Zaki and Meira.

References for "FP-Growth Algorithm"

1. Book: Han et al. *"Data Mining Concepts and Techniques"*, 3rd Edition, 2012, Morgan Kaufmann Publishers.
2. All text and images for some slides in "FP-Growth Algorithm" are based on the book by Han et al.

References for "Clustering"

1. Book: Steven S. Skiena, *"The Data Science Design Manual"*, 2017, Springer.
2. All text and images for some slides in "Clustering" are based on the book by Steven S. Skiena.

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Association Rules
   - Maximal and Closed Itemsets
   - Alternative Methods for Itemset Generation
     - FP-Growth Algorithm
     - ECLAT Algorithm
   - Interestingness Measures

3. Clustering
   - Introduction
   - K-Means Clustering
   - Hierarchical Clustering

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

Compact Representation of Frequent Itemsets

- Some itemsets are redundant because they have identical support as their supersets.
- Number of frequent itemsets = $3 \cdot \sum_{k=1}^{10} {}^{10}C_k$.
- Need a more compact representation.

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Maximal Frequent Itemsets

- An itemset is maximal frequent if it is frequent and none of its immediate supersets is frequent.

Figure 5.16. Maximal frequent itemset.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Maximal Frequent Itemsets

- What are the maximal frequent itemsets in this data.
- Minimum support threshold = 5.
- $(A1, A2, \ldots, A10)$, $(B1, B2, \ldots, B10)$, and $(C1, C2, \ldots, C10)$.

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Maximal Frequent Itemsets

Support threshold (by count) : 5
Frequent itemsets: {F}
Maximal itemsets: {F}

Support threshold (by count): 4
Frequent itemsets: {E}, {F}, {E,F}, {J}
Maximal itemsets: {E,F}, {J}

Support threshold (by count): 3
Frequent itemsets:
All subsets of {C,D,E,F} + {J}
Maximal itemsets:
{C,D,E,F}, {J}

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Closed Itemset and Closed Frequent Itemset

- An itemset X is closed if none of its immediate supersets has the same support as the itemset X.
- X is not closed if at least one of its immediate supersets has support count as X.
- An itemset is a closed frequent itemset if it is closed and its support is greater than or equal to minsup.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Closed Itemset and Closet Frequent Itemset

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%

Figure 5.17. An example of the closed frequent itemsets (with minimum support equal to 40%).

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Maximal Frequent and Closed Frequent Itemsets

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Frequent, Maximal Frequent, Closed Frequent Itemsets

- All maximal frequent itemsets are closed because none of the maximal frequent itemsets can have the same support count as their immediate supersets.

Figure 5.18. Relationships among frequent, closed, closed frequent, and maximal frequent itemsets.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice: general-to-specific or specific-to-general.

Figure 5.19. General-to-specific, specific-to-general, and bidirectional search.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice: equivalent classes.

(a) Prefix tree.

(b) Suffix tree.

Figure 5.20. Equivalence classes based on the prefix and suffix labels of itemsets.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice: breadth-first-search or depth-first-search.

(a) Breadth first (b) Depth first

Figure 5.21. Breadth-first and depth-first traversals.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice: general-to-specific or specific-to-general.

Figure 5.22. Generating candidate itemsets using the depth-first approach.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice: general-to-specific or specific-to-general.

Figure 5.22. Generating candidate itemsets using the depth-first approach.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Alternative Methods for Frequent Itemset Generation

- Database Representation: horizontal or vertical data layout.

Horizontal Data Layout

TID	Items
1	a,b,e
2	b,c,d
3	c,e
4	a,c,d
5	a,b,c,d
6	a,e
7	a,b
8	a,b,c
9	a,c,d
10	b

Vertical Data Layout

a	b	c	d	e
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

Figure 5.23. Horizontal and vertical data format.

Drawbacks of Apriori Algorithm

- Apriori candidate generate-and-test method significantly reduces the size of candidate sets, leading to good performance gain.
- However, it can suffer from two nontrivial costs:
  1. Need to generate a huge number of candidate sets.
     - For example, if there are 10^4 frequent 1-itemsets, the Apriori algorithm will need to generate more than 10^7 candidate 2-itemsets.
  2. Need to repeatedly scan the whole database and check a large set of candidates by pattern matching.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

FP-Growth Algorithm

- FP-Growth mines the complete set of frequent itemsets without such a costly candidate generation process.
- FP-growth, which adopts a divide-and-conquer strategy as follows:
  1. Compress the database representing frequent items into a frequent pattern tree, or FP-tree, which retains the itemset association information.
  2. Divides the compressed database into a set of conditional databases (projected trees) each associated with one "pattern fragment," and mines each database separately. For each "pattern fragment," only its associated data sets need to be examined.
- FP-Growth substantially reduce the size of the data sets to be searched, along with the "growth" of patterns being examined.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

FP-Growth Algorithm

1. Obtain the transaction database.

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

(a) Binary database

t	$\mathbf{i}(t)$
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

(b) Transaction database

x	A	B	C	D	E
$\mathbf{t}(x)$	1 3 4 5	1 2 3 4 5 6	2 4 5 6	1 3 5 6	1 2 3 4 5

(c) Vertical database

Figure 8.1. An example database.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm

1. Obtain the transaction database.
2. Sort the items in the transaction database by their support.

Item	Support
B	6
E	5
A	4
C	4
D	4

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm

1. Obtain the transaction database.
2. Sort the items in the transaction database by their support.
3. Sort the order of the items in each transaction of the database with descending order of support.

Item	Support
B	6
E	5
A	4
C	4
D	4

t	Transaction
1	BEAD
2	BEC
3	BEAD
4	BEAC
5	BEACD
6	BCD

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm

1. Obtain the transaction database.
2. Sort the items in the transaction database by their support.
3. Sort the order of the items in each transaction of the database with descending order of support.
4. Construct the FP-Tree step-by-step by adding each of the transactions (where items in them are reordered with descending support).

Item	Support
B	6
E	5
A	4
C	4
D	4

t	Transaction
1	BEAD
2	BEC
3	BEAD
4	BEAC
5	BEACD
6	BCD

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm

(a) $\langle 1, BEAD \rangle$ (b) $\langle 2, BEC \rangle$ (c) $\langle 3, BEAD \rangle$ (d) $\langle 4, BEAC \rangle$

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm

(e) $\langle 5, BEACD \rangle$ (f) $\langle 6, BCD \rangle$

Figure 8.7. Frequent pattern tree: bold edges indicate current transaction.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm

1. Obtain the transaction database.
2. Sort the items in the transaction database by their support.
3. Sort the order of the items in each transaction of the database with descending order of support.
4. Construct the FP-Tree step-by-step by adding each of the transactions (where items in them are reordered with descending support).
   - The FP-tree serves as an index in lieu of the original database.

(e) $\langle 5, BEACD \rangle$ (f) $\langle 6, BCD \rangle$

Figure 8.7. Frequent pattern tree: bold edges indicate current transaction.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm

1. Obtain the transaction database.
2. Sort the items in the transaction database by their support.
3. Sort the order of the items in each transaction of the database with descending order of support.
4. Construct the FP-Tree (call it R) step-by-step by adding each of the transactions (where items in them are reordered with descending support).
   - The FP-tree (R) serves as an index in lieu of the original database.
5. Given the FP-Tree (R), projected (conditional) FP-trees (databases) are built for each frequent item i in R in increasing order of support.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm — Projecting FP-Trees

1. To project R on item i , we find all the occurrences of i in the tree, and for each occurrence, we determine the corresponding path from the root to i (line 13).
2. The count of item i on a given path is recorded in $cnt(i)$ (line 14), and the path is inserted into the new projected tree R_X, where X is the itemset obtained by extending the prefix P with the item i.
3. While inserting the path, the count of each node in R_X along the given path is incremented by the path count $cnt(i)$.
4. We omit the item i from the path, as it is now part of the prefix.
5. The resulting FP-tree is a projection of the itemset X that comprises the current prefix extended with item i (line 9).
6. Call FPGrowth recursively with projected FP-tree R_X and the new prefix itemset X as the parameters (line 16).
7. The base case for the recursion happens when the input FP-tree R is a single path.
8. FP-trees that are paths are handled by enumerating all itemsets that are subsets of the path, with the support of each such itemset being given by the least frequent item in it (lines 2–6).

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm — Projecting FP-Trees

Algorithm 8.5: Algorithm FPGROWTH

```
// Initial Call:  R ← FP-tree(D),  P ← ∅,  F ← ∅
FPGROWTH (R, P, F, minsup):
Remove infrequent items from R
if ISPATH(R) then // insert subsets of  R into F
    foreach Y ⊆ R do
        X ← P ∪ Y
        sup(X) ← min_{x∈Y}{cnt(x)}
        F ← F ∪ {(X, sup(X))}
else  // process projected FP-trees for each frequent item  i
    foreach i ∈ R in increasing order of sup(i) do
        X ← P ∪ {i}
        sup(X) ← sup(i) // sum of  cnt(i)  for all nodes labeled  i
        F ← F ∪ {(X, sup(X))}
        R_X ← ∅ // projected FP-tree for  X
        foreach path ∈ PATHFROMROOT(i) do
            cnt(i) ← count of i in path
            Insert path, excluding i, into FP-tree R_X with count cnt(i)
        if R_X ≠ ∅ then FPGROWTH (R_X, X, F, minsup)
```

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm — Projecting FP-Trees

Figure 8.8. Projected frequent pattern tree for D.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm — Projecting FP-Trees

Figure 8.9. FPGrowth algorithm: frequent pattern tree projection.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

FP-Growth Algorithm — Mining Frequent Itemsets

- Let, $minsup = 3$.
- $P = D$: $\{DB(4), DE(3), DA(3), DBE(3), DBA(3), DEA(3), DBEA(3)\}$ (after removing C as it is infrequent — support = 2).
- $P = C$: $\{CB(4), CE(3), CBE(3)\}$.
- $P = A$: $\{AE(4), AB(4), AEB(4)\}$.
- $P = E$: $\{EB(5)\}$.

Figure 8.9. FPGrowth algorithm: frequent pattern tree projection.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

ECLAT Algorithm

- For each item, store a list of transaction ids (TIDS).

Horizontal Data Layout

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

Vertical Data Layout

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

↓

TID-list

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

ECLAT Algorithm

- Determine support of any k-itemset by intersection tid-lists of two of its $(k-1)$ subsets.
- 3 traversal approaches: top-down, bottom-up, and hybrid.
- Advantage: very fast support counting.
- Disadvantage: intermediate tid-lists may become too large for memory.

A
1
4
5
6
7
8
9

$\wedge$

B
1
2
5
7
8
10

$\rightarrow$

AB
1
5
7
8

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

Evaluation of Patterns and Rules

- Association rule algorithms often produce too many rules.
- Many of them are uninteresting or redundant.
- For example: $\{A, B, C\} \rightarrow \{D\}$ and $\{A, B\} \rightarrow \{D\}$ may convey redundant information and also have the same support and confidence
- "Interestingness Measures" can be used to prune and rank the resulting rules.
- We only used support and confidence measures.
- Other examples measures:

$$\text{Lift} = \frac{c(A \rightarrow B)}{s(B)} \tag{1}$$

All from: Kjetil Nørvåg's slides from previous year.

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Association Rules
   - Maximal and Closed Itemsets
   - Alternative Methods for Itemset Generation
     - FP-Growth Algorithm
     - ECLAT Algorithm
   - Interestingness Measures

3. Clustering
   - Introduction
   - K-Means Clustering
   - Hierarchical Clustering

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

1854 Broad Street Cholera Outbreak

- On 31 August 1854, a major outbreak of cholera occurred in Soho.
- Over the next three days, 127 people on or near Broad Street died. During the next week, three quarters of the residents had fled the area. By 10 September, 500 people had died and the mortality rate was 12.8 percent in some parts of the city. By the end of the outbreak, 616 people had died.
- Snow, the physician who eventually linked the outbreak to contaminated water, later called it "the most terrible outbreak of cholera which ever occurred in this kingdom."

All from: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak.

1854 Broad Street Cholera Outbreak

- By talking to local residents, Snow identified the source of the outbreak as the public water pump on Broad Street.
- Although Snow's chemical and microscope examination of a sample of the water from this Broad Street pump water did not conclusively prove its danger, his facts about the patterns of illness and death among residents in Soho persuaded the St James parish authorities to disable the well pump by removing its handle.

All from: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak.

1854 Broad Street Cholera Outbreak

- Snow later used a dot map to illustrate how cases of cholera occurred around this pump.

All from: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak.

1854 Broad Street Cholera Outbreak

- Snow later used a dot map to illustrate how cases of cholera occurred around this pump.

All from: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak.

Clustering — Toy Example

Clustering — Unsupervised Learning

- Consider the example observations from earlier.
- Given this set of data, how can we organize this data based on its attributes so that related observations are placed together and unrelated observations are excluded?

ID	Shape	Shape Color	Object_1	Color_1	Object_2	Color_2
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow

Example Clustering — Based on Shapes

Example Clustering — Based on Objects

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Application of Cluster Analysis

- Understanding: Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations.
- Summarization: Reduce the size of large data sets.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Notion of Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Types of Clustering

- A clustering is a set of clusters.
- Important distinction between hierarchical and partitional sets of clusters.
- Partitional Clustering: A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- Hierarchical Clustering: A set of nested clusters organized as a hierarchical tree.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Partitional Clustering

Original Points

A Partitional Clustering

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering

Hierarchical Clustering

Dendrogram

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Types of Clustering

- Exclusive versus Non-Exclusive:
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can belong to multiple classes or could be 'border' points.
- Fuzzy Clustering (one type of non-exclusive):
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1.
  - Weights must sum to 1.
  - Probabilistic clustering has similar characteristics.
- Partial versus Complete: In some cases, we only want to cluster some of the data.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Types of Clusters

- Well-separated clusters
- Prototype-based clusters
- Contiguity-based clusters
- Density-based clusters
- Described by an Objective Function

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Types of Clusters — Well-Seperated

- Well-Separated Clusters: A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Types of Clusters — Prototype-Based

- Prototype-Based:
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the prototype or "center" of a cluster, than to the center of any other cluster.
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster.

(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Types of Clusters — Contiguity-Based

- Contiguous Cluster (Nearest neighbor or Transitive): A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Types of Clusters — Density-Based

- Density-Based:
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.

(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Types of Clusters — Shared Property / Conceptual Clusters

- Finds clusters that share some common property or represent a particular concept.

(e) Conceptual clusters. Points in a cluster share some general property that derives from the entire set of points. (Points in the intersection of the circles belong to both.)

Figure 7.2. Different types of clusters as illustrated by sets of two-dimensional points.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Clustering Algorithms

- K-means and its variants.
- Hierarchical Clustering.
- Density-Based Clustering.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

K-Means Clustering

- Partitional clustering approach.
- Number of clusters, K, must be specified.
- Each cluster is associated with a centroid (center point).
- Each point is assigned to the cluster with the closest centroid.
- The basic algorithm is very simple.

Algorithm 7.1 Basic K-means algorithm.

```
Select K points as initial centroids.
repeat
    Form K clusters by assigning each point to its closest centroid.
    Recompute the centroid of each cluster.
until Centroids do not change.
```

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Example

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Example Iterations 1

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Example Iterations 2

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Example Iterations 3

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Example Iterations 4

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Example Iterations 5

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Example Iterations 6

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means Clustering — Details

- Simple iterative algorithm.
- Initial centroids are often chosen randomly. Therefore, clusters produced can vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster, but other definitions are possible (based on different similarity functions).
- K-means will converge for common proximity measures with appropriately defined centroid (based on different similarity functions).
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to: until relatively few points change clusters.
- Time complexity is: $\mathcal{O}(n \cdot K \cdot I \cdot d)$.
- Space complexity is: $\mathcal{O}((n + K) \cdot d)$.
  - n = number of points.
  - K = number of clusters.
  - I = number of iterations.
  - d = number of attributes.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Evaluating K-Means Clusters

- Most common measure is Sum of Squared Error (SSE).
  - For each point, the error is the distance to the nearest cluster center.
  - To get SSE, we square these errors and sum them.

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \Big(\text{dist}(m_i, x)\Big)^2. \tag{2}$$

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i.
  - We can show that m_i corresponds to the center (mean) of the cluster:

$$c_i = \frac{1}{n_i} \sum_{x \in C_i} x. \tag{3}$$

  - where, n_i is the number of objects in the ith cluster.
- One easy way to reduce SSE is to increase K, the number of clusters.
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Parameters — How to Select K in K-Means

- We seek from error curve: value K where the rate of decline decreases.
- Error curve should look something like an arm in typing position: it slopes down rapidly from shoulder to elbow, and then slower from the elbow to the wrist.
- Want K to be located exactly at the elbow.
- Easier to identify when compared to a similar MSE error plot for random centers, since the relative rate of error reduction for random centers should be analogous to what we see past the elbow. The slow downward drift is telling us the extra clusters are not doing anything special for us.

All from: Steven S. Skiena, "The Data Science Design Manual", 2017, Springer.

Parameters — How to Select k in K-Means

Figure 10.16: The error curve for k-means clustering on the point set of Figure 10.12, showing a bend in the elbow reflecting the three major clusters in the data. The error curve for random cluster centers is shown for comparison.

All from: Steven S. Skiena, "The Data Science Design Manual", 2017, Springer.

Parameters: Selecting Initial Centroids

Original Points

Optimal Clustering

Sub-optimal Clustering

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Two Different K-Means Clusterings

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Solutions to Initial Centroids Problem

1. Multiple runs: helps, but probability is not on your side.
2. Sample and use hierarchical clustering to determine initial centroids.
3. Select more than K initial centroids and then select among these initial centroids. Select most widely separated.
4. K-Means++.
5. Bisecting K-means.
   - Not as susceptible to initialization issues.
6. Post-processing.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

K-Means++: An Approach to Selecting Initial Centroids

- This approach can be slower than random initialization, but very consistently produces better results in terms of SSE.
- Algorithm to select initial centroids C:
  1. Select an initial point at random to be the first centroid.
  2. For $k-1$ steps:
     1. For each of the N points x_i, $1 \leq i \leq N$, find the minimum squared distance to the currently selected centroids, $C_1, \ldots, C_j$, $1 \leq j < k$, i.e., $\min_j(d(C_j, x_i))^2$.
     2. Randomly select a new centroid by choosing a point with probability proportional to $\frac{\min_j(d(C_j,x_i))^2}{(\sum_i d(C_j,x_i))^2}$.
  3. End For.
- Afterwards, continue K-Means as usual.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Bisecting K-Means

- Bisecting K-means algorithm: variant of K-means that can produce a partitional or a hierarchical clustering.
- Straightforward extension of the basic K-means algorithm:
  1. To obtain K clusters, split the set of all points into two clusters.
  2. Select one of these clusters to split, and so on,
     until K clusters have been produced.

Algorithm 7.3 Bisecting K-means algorithm.

```
Initialize the list of clusters to contain the cluster consisting of all points.
repeat
    Remove a cluster from the list of clusters.
    {Perform several "trial" bisections of the chosen cluster.}
    for i = 1 to number of trials do
        Bisect the selected cluster using basic K-means.
    end for
    Select the two clusters from the bisection with the lowest total SSE.
    Add these two clusters to the list of clusters.
until The list of clusters contains K clusters.
```

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Bisecting K-Means

(a) Iteration 1. (b) Iteration 2. (c) Iteration 3.

Figure 7.8. Bisecting K-means on the four clusters example.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Pre-processing and Post-processing

- Pre-processing:
  - Normalize the data.
  - Eliminate outliers.
- Post-processing:
  - Eliminate small clusters that may represent outliers.
  - Split 'loose' clusters, i.e., clusters with relatively high SSE.
  - Merge clusters that are 'close' and that have relatively low SSE.
  - Can use these steps during the clustering process.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Limitations of K-Means

- K-means has problems when clusters are of differing:
  - Sizes
  - Densities
  - Non-globular shapes
- K-means has problems when the data contains outliers.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Limitations of K-Means: Differing Sizes

(a) Original points. (b) Three K-means clusters.

Figure 7.9. K-means with clusters of different size.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Limitations of K-Means: Differing Densitites

(a) Original points.

(b) Three K-means clusters.

Figure 7.10. K-means with clusters of different density.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Limitations of K-Means: Non-Globular Clusters

(a) Original points. (b) Two K-means clusters.

Figure 7.11. K-means with non-globular clusters.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Association Rules
- Maximal and Closed Itemsets
- Alternative Methods for Itemset Generation
  - FP-Growth Algorithm
  - ECLAT Algorithm
- Interestingness Measures

3 Clustering
- Introduction
- K-Means Clustering
- Hierarchical Clustering

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree.
- Can be visualized as a dendrogram: a tree like diagram that records the sequences of merges or splits.

(a) Dendrogram. (b) Nested cluster diagram.

Figure 7.13. A hierarchical clustering of four points shown as a dendrogram and as nested clusters.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters.
- Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level.
- They may correspond to meaningful taxonomies.
- For example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, ...).

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering

- Two main types of hierarchical clustering:
  1. Agglomerative:
     - Start with the points as individual clusters.
     - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left.
  2. Divisive:
     - Start with one, all-inclusive cluster.
     - At each step, split a cluster until each cluster contains a point (or there are k clusters).
- Traditional hierarchical algorithms use a similarity or distance matrix.
  - Merge or split one cluster at a time.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Agglomerative Clustering Algorithm

- Most popular hierarchical clustering technique.
- Basic algorithm is straightforward:
  1. Compute the proximity matrix.
  2. Let each data point be a cluster.
  3. Repeat
     1. Merge the two closest clusters.
     2. Update the proximity matrix.
  4. Until only a single cluster remains.
- Key operation is the computation of the proximity of two clusters.
- Different approaches to defining the distance between clusters distinguish the different algorithms.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Starting Situation

- Start with clusters of individual points and a proximity matrix.

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Intermediate Situation

- After some merging steps, we have some clusters.

	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.

	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

After Merging

- The question is: "How do we update the proximity matrix?".

	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Define Inter-Cluster Distance

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Define Inter-Cluster Distance

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Define Inter-Cluster Distance

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Define Inter-Cluster Distance

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Define Inter-Cluster Distance

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: MIN or Single Link

- Proximity of two clusters is based on the two closest points in the different clusters.
- Determined by one pair of points, i.e., by one link in the proximity graph.
- Example:

Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: MIN

Nested Clusters

Dendrogram

$$
\begin{aligned}
dist(\{3,6\},\{2,5\}) &= \min(dist(3,2), dist(6,2), dist(3,5), dist(6,5)) \\
&= \min(0.15, 0.25, 0.28, 0.39) \\
&= 0.15.
\end{aligned}
$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Strengths of MIN

- Can handle non-elliptical shapes.

Original Points

Six Clusters

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Weakness of MIN

- Sensitive to noise and outliers.

Original Points

Two Clusters

Three Clusters

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: MAX or Complete Linkage

- Proximity of two cluster is based on the two most distant points in the different clusters.
  - Determined by all pairs of points in the two clusters.

Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: MAX or Complete Linkage

Nested Clusters

Dendrogram

$$
\begin{aligned}
dist(\{3,6\},\{4\}) &= \max(dist(3,4), dist(6,4)) \\
&= \max(0.15, 0.22) \\
&= 0.22. \\
dist(\{3,6\},\{2,5\}) &= \max(dist(3,2), dist(6,2), dist(3,5), dist(6,5)) \\
&= \max(0.15, 0.25, 0.28, 0.39) \\
&= 0.39. \\
dist(\{3,6\},\{1\}) &= \max(dist(3,1), dist(6,1)) \\
&= \max(0.22, 0.23) \\
&= 0.23.
\end{aligned}
$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Strengths of MAX

- Less susceptible to noise and outliers.

Original Points **Two Clusters**

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Limitations of MAX

- Tends to break large clusters.
- Biased towards globular clusters.

Original Points

Two Clusters

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\text{proximity}(\text{cluster}_1, \text{cluster}_2) = \frac{\sum_{\substack{p_i \in \text{cluster}_i \\ p_j \in \text{cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{cluster}_i| \cdot |\text{cluster}_j|}. \tag{4}$$

- Need to use average connectivity for scalability since total proximity favors large clusters.

Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Group Average

(a) Group average clustering.

(b) Group average dendrogram.

$$
\begin{aligned}
dist(\{3,6,4\},\{1\}) &= (0.22+0.37+0.23)/(3\times 1) \\
&= 0.28 \\
dist(\{2,5\},\{1\}) &= (0.24+0.34)/(2\times 1) \\
&= 0.29 \\
dist(\{3,6,4\},\{2,5\}) &= (0.15+0.28+0.25+0.39+0.20+0.29)/(3\times 2) \\
&= 0.26
\end{aligned}
$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Group Average

- Compromise between single and complete link.
- Strengths: less susceptible to noise and outliers.
- Limitations: biased towards globular clusters.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: time and Space Requirements

- $\mathcal{O}(N^2)$ space since it uses the proximity matrix.
  - N is the number of points.
- $\mathcal{O}(N^3)$ time in many cases.
  - There are N steps and at each step the size, N^2, proximity matrix must be updated and searched.
  - Complexity can be reduced to $\mathcal{O}(N^2 \log(N))$ time with some cleverness.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone.
- No global objective function is directly minimized.
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers.
  - Difficulty handling clusters of different sizes and non-globular shapes.
  - Breaking large clusters.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.