Data Warehouse and Data Mining

Dhruv Gupta

dhruv.gupta@ntnu.no

26-January-2021

NTNU | Norwegian University of Science and Technology

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

2 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

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

2 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Administrative

1. First Assignment
   - Available and due by 26.February.2021 (date has changed).
2. Assignment Groups: groups of up to three persons allowed.
3. Guest lecture by BearingPoint confirmed!
   - Mark your calendars: March 2, 2021 at 0815 - 1000 HRS.
4. TA Consultation Hours
   - Shiva Shadrooh (*shiva.shadrooh@ntnu.no*) : (Fridays 1400-1600 HRS).
   - Sjur Wold (*sjurwo@stud.ntnu.no*) : (Tuesdays 1400-1600 HRS).
   - Patrik Kjærran (*patrikkj@stud.ntnu.no*): (Thursdays 1400-1600 HRS).
5. Volunteers for feedback regarding course
   - Interested? Please contact me by email!

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

2 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

References for "Data" and "Association Rules"

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

References for "Data" and "Association Rules"

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 "Data" and "Association Rules" subsection are based on the book by Zaki and Meira.

References for "Data" and "Association Rules"

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 "Data" and "Association Rules" are based on the book by Han et al.

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

2 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

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

2 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Curse of Dimensionality

The term curse of dimensionality was coined by Bellman (1961) to indicate that the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with the number of variables, that is, with the dimensionality of the function.

All from: Cichocki et al. "Tensor Networks for Dimensionality Reduction and Large-Scale Optimization Part 1 Low-Rank Tensor Decompositions", Vol. 9 No. 4–5 (2016), Foundations and Trends in Machine Learning.

Curse of Dimensionality

Table 1.1. Extract from the Iris dataset

	Sepal length X_1	**Sepal width** X_2	**Petal length** X_3	**Petal width** X_4	**Class** X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

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

Curse of Dimensionality

Figure 6.2. 3-dimensional Iris data hyperspace with a 2-dimensional hyperplane. Points in white are above the plane, whereas points in gray are below the plane.

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

Curse of Dimensionality

Figure 6.1. Iris data hyperspace: hypercube (solid; with $l = 4.12$) and hypersphere (dashed; with $r = 2.19$).

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

Curse of Dimensionality

Figure 6.3. Volume of a unit hypersphere.

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

Curse of Dimensionality

Figure 6.4. Hypersphere inscribed inside a hypercube: in (a) two and (b) three dimensions.

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

Curse of Dimensionality

Figure 6.5. Conceptual view of high-dimensional space: (a) two, (b) three, (c) four, and (d) higher dimensions. In d dimensions there are 2^d "corners" and 2^{d-1} diagonals. The radius of the inscribed circle accurately reflects the difference between the volume of the hypercube and the inscribed hypersphere in d dimensions.

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

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.
- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful.

Randomly generate 500 points

Compute difference between max and min distance between any pair of points

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

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality.
  - Reduce amount of time and memory required by data mining algorithms.
  - Allow data to be more easily visualized.
  - May help to eliminate irrelevant features or reduce noise.
- Techniques:
  - Principal Components Analysis (PCA).
  - Singular Value Decomposition.
  - Others: supervised and non-linear techniques.

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

Principal Component Analysis (PCA)

- Principal Components Analysis (PCA) is a linear algebra technique for continuous attributes that finds new attributes (principal components) that are:
  1. are linear combinations of the original attributes,
  2. are orthogonal (perpendicular) to each other, and
  3. capture the maximum amount of variation in the data.

Figure 7.2. Best one-dimensional or line approximation.

All from: 1. Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pea
2. Zaki and Meira, "Data Mining and Machine Learning: Fundamental Co
Press.

Principal Component Analysis (PCA)

(a) Optimal basis (b) Nonoptimal basis

Figure 7.3. Best two-dimensional approximation.

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

Linear Discriminant Analysis (LDA)

Figure 4.28: Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. Informally, multiple discriminant methods seek the optimum such subspace, i.e., the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{w}_1$.

All from: Duda et al. "Pattern Classification", 2nd Edition, 2001, John Wiley & Sons, Inc.

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

2 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Feature Selection

- Another way to reduce dimensionality of data.
- Redundant features:
  - Duplicate much or all of the information contained in one or more other attributes.
  - Example: purchase price of a product and the amount of sales tax paid.
- Irrelevant features:
  - Contain no information that is useful for the data mining task at hand.
  - Example: students' ID is often irrelevant to the task of predicting students' GPA.
  - Many techniques developed, especially for classification.

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

Feature Selection — Ideal Approach

- Irrelevant and redundant attributes can be eliminated by using common sense domain knowledge.
- A more systematic approach is however required.
- Ideal Approach:
  - Enumerate all possible subsets of features.
  - For each subset obtain performance results form the data mining algorithm being used.
- Pros: reflects the objective and bias of the data mining algorithm that will eventually be used.
- Cons: number of subsets involving n attributes is 2^n.

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

Feature Selection — Other Approaches

- Embedded Approaches: algorithm itself decides which attributes to use and which to ignore (e.g., decision trees).
- Filter Approaches: features are selected independently and separately from the algorithm being run (e.g., where pairwise correlation is as low as possible).
- Wrapper Approaches: same as Ideal Approach but without enumerating all possible subsets.

Figure 2.11. Flowchart of a feature subset selection process.

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes.
- Three general methodologies:
  - Feature extraction.
    - Example: extracting edges from images.
  - Feature construction.
    - Example: dividing mass by volume to get density.
  - Mapping data to new space.
    - Example: Fourier and wavelet analysis.

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

Feature Creation

- Example of mapping data to new space via Fourier Transform.

Two Sine Waves + Noise

Frequency

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Discretization

- Certain classification algorithms, require data in the form of categorical attributes.
- Algorithms that find association patterns require that the data be in the form of binary attributes.
- Discretization: transform a continuous attribute into a categorical attribute.
- Binarization: Both continuous and discrete attributes may need to be transformed into one or more binary attributes.

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

Binarization

- A simple approach to binarize m categorical values:
  1. Assign a unique integer in the interval $[0, m-1]$. Maintain order during assignment if the attribute is ordinal.
  2. Convert each of these m integers to a binary number.
  3. $n = \lceil \log_2(m) \rceil$ binary attributes are required to represent these integers.
- Problems: unintended correlations introduced.

Categorical Value	**Integer Value**	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
okay	2	0	1	0
good	3	0	1	1
great	4	1	0	0

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

Binarization

- For association problems asymmetrical attributes are required.
- Therefore necessary to introduce one binary attribute for each categorical value.

Categorical Value	**Integer Value**	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
okay	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

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

Discretization

- Transformation of a continuous attribute to a categorical attribute involves two subtasks:
  1. First step: continuous attribute are sorted and then are divided into n intervals by specifying $n-1$ split points.
  2. Second step: all the values in one interval are mapped to the same categorical value.
- Unsupervised Approaches: label associated with the objects are not used.
  - Example: equal width, equal frequency / equal depth, and k-means.
- Supervised Approaches: labels associated with the objects are used.
  - Example: a simple approach for partitioning a continuous attribute starts by bisecting the initial values so that the resulting two intervals give minimum entropy.

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

Discretization

Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

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

Discretization

Equal interval width approach used to obtain 4 values.

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

Discretization

Equal frequency approach used to obtain 4 values.

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

Discretization

K-means approach to obtain 4 values.

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Simple Functions

- Attribute Transform: a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.
- Using Simple Functions:
  - Using transformations such as x^k, $\log(x)$, e^x, $\sqrt{x}$, $1/x$, or $|x|$.
  - Variable transformations should be applied with caution since they change the nature of the data.
  - Does the transformation apply to all values?
  - Especially negative values and 0?
  - What is the effect of the transformation on the values between 0 and 1?

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

Normalization or Standardization

- Normalization refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, and range.
- Take out unwanted, common signal, e.g., seasonality.
- In statistics, standardization refers to subtracting off the means and dividing by the standard deviation.

$$x' = \frac{x - \bar{x}}{s_x} \tag{1}$$

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Similarity and Dissimilarity

- Similarity between two objects:
  - A numerical measure indicating degree to which the two objects are alike.
  - Usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).
- Dissimilarity between two objects:
  - A numerical measure of the degree to which the two objects are different.
  - Dissimilarities are lower for more similar pairs of objects.
  - Dissimilarities sometimes fall in the interval [0,1], but it is also common for them to range from 0 to ∞.
  - Distances, which are dissimilarities with certain properties.

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

Similarity and Dissimilarity for Simple Attributes

Attribute Type	**Dissimilarity**	**Similarity**
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min_d}{max_d - min_d}$

Figure: The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.

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

Distances (Dissimilarities with Certain Properties)

- Euclidean Distance:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{2}$$

- where, n is the number of dimensions (attributes) and x_k and y_k are respectively the kth attributes (components) or data objects x and y.
- Standardization is necessary if scales differ.

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

Euclidean Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

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

Minkowski Distance

- Minkowski Distance — generalization of Euclidean distance:

$$d(x, y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r} \tag{3}$$

- where, r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the kth attributes (components) or data objects x and y.

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

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L 1 norm) distance.
  - A common example of this for binary vectors is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance.
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm, Uniform norm, and Chebyshev) distance.
  - This is the maximum difference between any component of the vectors.
  - That is,

$$d(x, y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r} = \max_{k}^{n} |x_k - y_k|. \tag{4}$$

- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

All from: 1. Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.
2. Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

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

Common Properties of Distance

- Distances, such as the Euclidean distance, have some well known properties.
  - Positivity: $d(x, y) \geq 0$ for all x and y and $d(x, y) = 0$ if and only if $x = y$.
  - Symmetry: $d(x, y) = d(y, x)$ for all x and y.
  - Triangle Inequality: $d(x, y) \leq d(x, y) + d(y, z)$ for all points x, y, and z.
- A distance that satisfies these properties is a metric.

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

Similarity between Binary Vectors

- Common situation is that objects, x and y, have only binary attributes.
- Compute similarities using the following quantities:
  - f_{00} = the number of attributes where x was 0 and y was 0.
  - f_{01} = the number of attributes where x was 0 and y was 1.
  - f_{10} = the number of attributes where x was 1 and y was 0.
  - f_{11} = the number of attributes where x was 1 and y was 1.
- Simple Matching Coefficient:

$$SMC = \frac{\text{number of matches}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}}. \tag{5}$$

- Jaccard Coefficient:

$$J = \frac{\text{number of 11 matches}}{\text{number of non-zero attributes}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}. \tag{6}$$

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

Simple Matching and Jaccard Coefficient Example

- $x = \langle 1, 0, 0, 0, 0, 0, 0, 0, 0, 0 \rangle$
- $y = \langle 0, 0, 0, 0, 0, 0, 1, 0, 0, 1 \rangle$
  - $f_{00} = 7$ the number of attributes where x was 0 and y was 0.
  - $f_{01} = 2$ the number of attributes where x was 0 and y was 1.
  - $f_{10} = 1$ the number of attributes where x was 1 and y was 0.
  - $f_{11} = 0$ the number of attributes where x was 1 and y was 1.
- Simple Matching Coefficient:

$$SMC = \frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{0 + 7}{7 + 2 + 1 + 0} = \frac{7}{10}. \tag{7}$$

- Jaccard Coefficient:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2 + 1 + 0} = 0. \tag{8}$$

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

Recall — Document Data

- Each document becomes a term vector.
- Each term is a component (attribute) of the vector.
- The value of each component is the number of times the corresponding term occurs in the document.
- Example of sparse data matrix.

	$\mathbf{word}_1$	$\mathbf{word}_2$	$\mathbf{word}_3$	$\ldots$	$\mathbf{word}_{\lvert\mathcal{V}\rvert}$
$\mathbf{document}_1$	10	0	0	$\ldots$	5
$\mathbf{document}_2$	1	2	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
$\mathbf{document}_n$	0	1	0	$\ldots$	0

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

Cosine Similarity

- If x and y are two document vectors, then:

$$\cos(x, y) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} = \frac{\sum_{k=1}^{n} x_k \cdot y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \cdot \sqrt{\sum_{k=1}^{n} y_k^2}}. \tag{9}$$

- where, $\langle x, y \rangle$ indicates the inner product or vector dot product of vectors, x and y, and $\|x\|$ is the length of vector x.
- Example:

$$d_1 = \langle 3, 2, 0, 5, 0, 0, 0, 2, 0, 0 \rangle,$$
$$d_2 = \langle 1, 0, 0, 0, 0, 0, 0, 1, 0, 2 \rangle.$$

$$\langle d_1, d_2 \rangle = 3 \cdot 1 + 2 \cdot 0 + 0 \cdot 0 + 5 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 2 = 5$$

$$\|d_1\| = \sqrt{(3 \cdot 3 + 2 \cdot 2 + 0 \cdot 0 + 5 \cdot 5 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 2 + 0 \cdot 0 + 0 \cdot 0)} = \sqrt{42} = 6.481$$

$$\|d_2\| = \sqrt{(1 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 2 \cdot 2)} = \sqrt{6} = 2.449$$

$$\cos(d_1, d_2) = \frac{5}{6.481 \cdot 2.449} = 0.3150$$

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Correlation

- Correlation between two data objects that have binary or continuous variables is a measure of the linear relationship between the attributes of the objects.
- Perfect Correlation: correlation is always in the range -1 to 1. A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship.
- That is, $x_k = a \cdot y_k + b$, where a, and b are constants.
- Example 1:

$$x = \langle -3, 6, 0, 3, -6 \rangle \text{ and } y = \langle 1, -2, 0, -1, 2 \rangle.$$

- where, $x = -3 \cdot y$ implies correlation is -1.
- Example 2:

$$x = \langle 3, 6, 0, 3, 6 \rangle \text{ and } y = \langle 1, 2, 0, 1, 2 \rangle.$$

- where, $x = 3 \cdot y$ implies correlation is 1.

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

Pearson's Correlation Coefficient

- Pearson's Correlation Coefficient between two data objects, x and y is defined by the following equation:

$$\text{corr}(x, y) = \frac{\text{covariance(x,y)}}{\text{std_dev}(x) \cdot \text{std_dev}(y)} = \frac{s_{xy}}{s_x \cdot s_y}. \tag{10}$$

- where, the standard statistical notation and definitions are:

$$\text{covariance}(x, y) = s_{xy} = \frac{1}{n-1} \cdot \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y}). \tag{11}$$

$$\text{std_dev}(x) = s_x = \sqrt{\frac{1}{n-1} \cdot \sum_{k=1}^{n} (x_k - \bar{x})^2}. \tag{12}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } x. \tag{13}$$

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

Visualizing Correlations

Scatter plots showing the similarity from –1 to 1.

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

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

2 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Association Rule Mining

1. Association Analysis: discovering interesting relationships hidden in large data sets.
2. Uncovered relationships can be represented in the form of association rules or sets of frequent items.
3. When applied to retail transactions, commonly known as **Market Basket Analysis**.
4. Other application areas: medical diagnosis, Web mining, and scientific data analysis.

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

Association Rule Mining

1. Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.
2. Example of association rules:
   - $\{Diaper\} \rightarrow \{Beer\}$
   - $\{Milk, Bread\} \rightarrow \{Eggs, Cola\}$
   - $\{Beer, Bread\} \rightarrow \{Milk\}$
3. Note that implication means co-occurence, not causality.

T-ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Cola
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Cola

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Definitions

- Binary Representation: presence of an item in a transaction is often considered more important than its absence, therefore an item is represented as an asymmetric binary variable.

T-ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Cola
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Cola

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

Definitions

- Binary Representation: presence of an item in a transaction is often considered more important than its absence, therefore an item is represented as an asymmetric binary variable.

T-ID	Bread	Milk	Diapers	Beers	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

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

Definitions

- Set of all transactions: $T = \{t_1, t_2, \ldots, t_N\}$.
- Set of all Items: $I = \{i_1, i_2, \ldots, i_d\}$.
- Each transaction t_i contains a subset of items chosen from I.
- Itemset: a collection of zero or more items.
- $k-$Itemset: itemset containing k items.
  Example: $\{Beer, Diapers, Milk\}$ is a 3-itemset.

T-ID	Bread	Milk	Diapers	Beers	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

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

Definitions

- Support Count ($\sigma(X)$): number of transactions that contain a particular itemset.

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}| \tag{14}$$

- Example: $\sigma(\{Beer, Diapers, Milk\}) = 2$.

T-ID	Bread	Milk	Diapers	Beers	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

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

Definitions

- Association Rule: is an implication expression of the form $X \rightarrow Y$, where, $X \cap Y = \emptyset$.
- Support $s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$.
- Example: $s(\{Milk, Diaper\} \rightarrow \{Beer\}) = \frac{\sigma(\{Milk, Diaper, Beer\})}{|T|} = 2/5 = 0.4$
- Confidence $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$.
- Example: $c(\{Milk, Diaper\} \rightarrow \{Beer\}) = \frac{\sigma(\{Milk, Diaper, Beer\})}{\sigma(\{Milk, Diaper\})} = 2/3 = 0.\bar{6}$

T-ID	Bread	Milk	Diapers	Beers	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Association Rule Mining Problem

- Given a set of transactions T, the goal of association rule mining is to find all rules having:
  - support $\geq$ minimum support threshold,
  - confidence $\geq$ minimum confidence threshold.
- Minimum support (*minsup*) and confidence (*minconf*) represent thresholds on interesting association rules.
- Brute-Force Approach:
  1. List all possible association rules.
  2. Compute the support and confidence for each rule.
  3. Prune rules that do not satisfy the minsup and minconf thresholds.
- Brute-force is computationally prohibitive!

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

Brute-Force Association Rule Mining

- Consider, we have d unique items.
- Total number of itemsets = 2^d.
- Total number of possible association rules

$$R = \sum_{k=1}^{d} \left[{}^{d}C_k \cdot \sum_{j=1}^{d-k} {}^{d-k}C_j \right],$$
$$= 3^d - 2^d + 1.$$

- Example: for $d = 6$, then $R = 602$ rules to analyze.

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

Mining Association Rules

- Consider, all rules that can be generated from the itemset $\{Milk, Diaper, Beer\}$:
  1. $\{Milk, Diaper\} \rightarrow \{Beer\}$ $(s = 0.4, c = 0.\bar{6})$.
  2. $\{Milk, Beer\} \rightarrow \{Diaper\}$ $(s = 0.4, c = 1.0)$.
  3. $\{Diaper, Beer\} \rightarrow \{Milk\}$ $(s = 0.4, c = 0.\bar{6})$.
  4. $\{Beer\} \rightarrow \{Milk, Diaper\}$ $(s = 0.4, c = 0.\bar{6})$.
  5. $\{Diaper\} \rightarrow \{Milk, Beer\}$ $(s = 0.4, c = 0.5)$.
  6. $\{Milk\} \rightarrow \{Diaper, Beer\}$ $(s = 0.4, c = 0.5)$.

T-ID	Bread	Milk	Diapers	Beers	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

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

Mining Association Rules

- Consider, all rules that can be generated from the itemset $\{Milk, Diaper, Beer\}$:
  1. $\{Milk, Diaper\} \rightarrow \{Beer\}$ $(s = 0.4, c = 0.\bar{6})$.
  2. $\{Milk, Beer\} \rightarrow \{Diaper\}$ $(s = 0.4, c = 1.0)$.
  3. $\{Diaper, Beer\} \rightarrow \{Milk\}$ $(s = 0.4, c = 0.\bar{6})$.
  4. $\{Beer\} \rightarrow \{Milk, Diaper\}$ $(s = 0.4, c = 0.\bar{6})$.
  5. $\{Diaper\} \rightarrow \{Milk, Beer\}$ $(s = 0.4, c = 0.5)$.
  6. $\{Milk\} \rightarrow \{Diaper, Beer\}$ $(s = 0.4, c = 0.5)$.
- Rules originating from the same itemset have identical support but can have different confidence.
- Thus, we may decouple the support and confidence requirements.

Mining Association Rules

- Two-step approach:
  1. Frequent Itemset Generation: Generate all itemsets whose support $\geq$ minimum support.
  2. Rule Generation: Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset.
- Frequent itemset generation is still a computationally expensive process.

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

Frequent Itemset Generation — Brute-Force Approach

- Given d items, there are 2^d possible candidate itemsets.

Figure 8.2. Itemset lattice and prefix-based search tree (in bold).

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

Brute-Force Frequent Itemset Generation

- Each itemset in the lattice is a candidate frequent itemset.
- Count the support of each candidate by scanning the database.
- Match each transaction against every candidate.
- Complexity — $\mathcal{O}(NMw)$ — very expensive since $M = 2^d$.

Transaction Database

T-ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Cola
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Cola

N

w (width of transaction)

List of Candidates

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

Improving Brute-Force Frequent Itemset Generation

- Reduce the number of candidates (M)
  - Complete search: $M = 2^d$.
  - Use pruning techniques to reduce M.
- Reduce the number of comparisons (NM)
  - Use efficient data structures to store the candidates or transactions.
  - No need to match every candidate against every transaction.

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 Data
- Curse of Dimensionality
- Feature Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformations
- Similarity and Dissimilarity
- Correlation

3 Association Rules
- Association Rule Mining
- Definitions
- Problem Definition

4 Summary

Summary — Feature Selection — Other Approaches

- Embedded Approaches: algorithm itself decides which attributes to use and which to ignore (e.g., decision trees).
- Filter Approaches: features are selected independently and separately from the algorithm being run (e.g., where pairwise correlation is as low as possible).
- Wrapper Approaches: same as Ideal Approach but without enumerating all possible subsets.

Figure 2.11. Flowchart of a feature subset selection process.

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

Summary — Discretization

Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

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

Summary — Euclidean Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

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

Summary — Frequent Itemset Generation

- Given d items, there are 2^d possible candidate itemsets.

Figure 8.2. Itemset lattice and prefix-based search tree (in bold).

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

Summary — Administrative

1. First Assignment
   - Available and due by 26.February.2021 (date has changed).
2. Assignment Groups: groups of up to three persons allowed.
3. Guest lecture by BearingPoint confirmed!
   - Mark your calendars: March 2, 2021 at 0815 - 1000 HRS.
4. TA Consultation Hours
   - Shiva Shadrooh (*shiva.shadrooh@ntnu.no*) : (Fridays 1400-1600 HRS).
   - Sjur Wold (*sjurwo@stud.ntnu.no*) : (Tuesdays 1400-1600 HRS).
   - Patrik Kjærran (*patrikkj@stud.ntnu.no*): (Thursdays 1400-1600 HRS).
5. Volunteers for feedback regarding course
   - Interested? Please contact me by email!