University of Southern Denmark
IMADA

DM566: Data Mining and Machine Learning

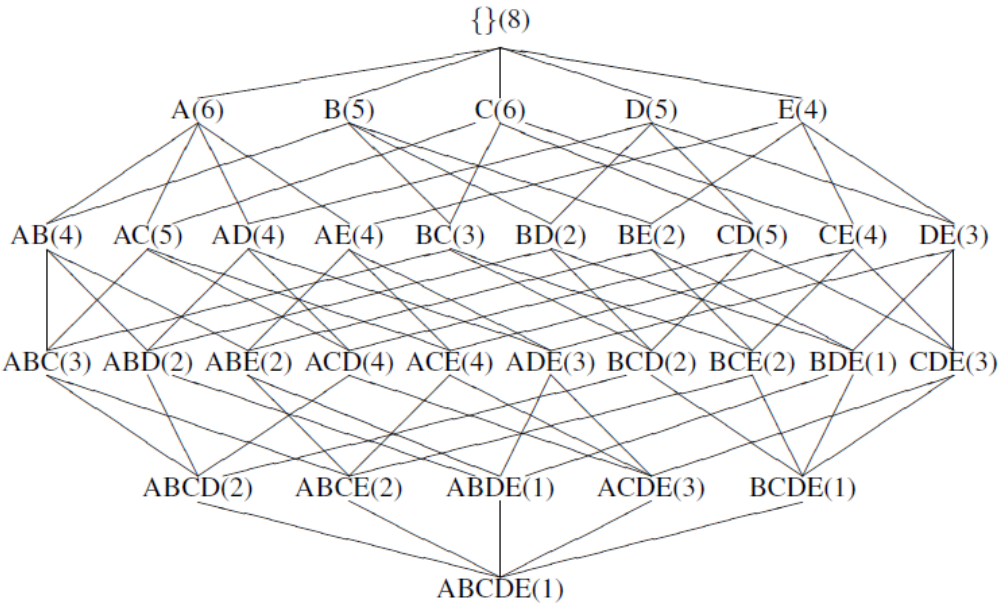
Spring term 2022

Exercise 3: Closed Frequent Itemsets, Apriori, Color Histrograms

Exercise 3-1 Support based on closed frequent itemsets (1 point)

1. We are given the following database of transactions. The support of all itemsets are computed below.

TID	A	B	C	D	E
1	0	1	0	0	0
2	1	0	1	1	1
3	1	1	1	0	1
4	0	0	1	1	0
5	1	1	1	1	1
6	1	0	1	1	1
7	1	1	0	0	0
8	1	1	1	1	0



Identify the closed frequent itemsets for the support thresholds $\sigma = 4$ and $\sigma = 2$, respectively. Which are they, and what do you observe?

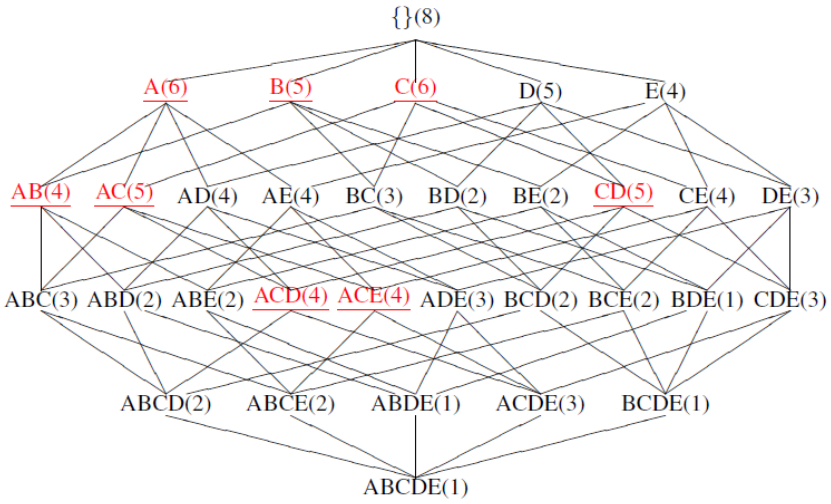
Suggested solution:

Note that

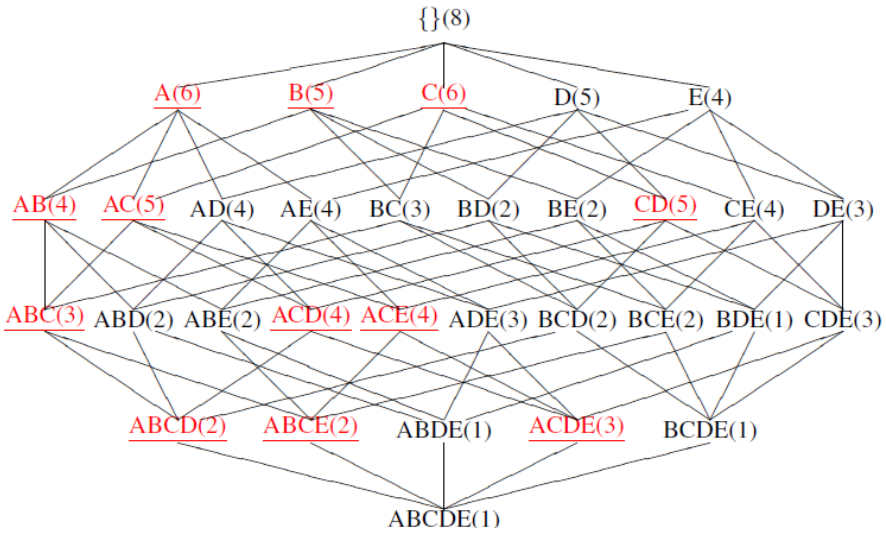
An itemset is frequent if its support is $\geq$ minsup.

An itemset is closed if it has no superset with the same frequency.

cfi for $\sigma = 4$:



cfi for $\sigma = 2$:



Observation: the closed frequent itemsets for $\sigma = 4$ are a subset of the cfi for $\sigma = 2$.

2. Sketch an algorithm (pseudo code) to find the support for all frequent itemsets, using only the set of closed frequent itemsets as information.

   Suggested solution:
   Let's say that you have the set of all closed frequent itemsets C and that you want to know the support of an itemset F.

   What you need to do is very simple. You need to compare F with all the frequent closed itemsets. You have to find the smallest closed itemset W such that F is included in W. Then the support of F is the support of W.

Exercise 3-2 Apriori (1 point)

Consider the following transaction database D over the items $I = \{A, B, C, D, E, F\}$.

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

Given the support threshold $\sigma = 2$, apply the Apriori algorithm and extract all frequent itemsets w.r.t. the given threshold. Include all the steps that you followed, particularly the candidate set C_k before and after pruning. Also give explicitly the solution of frequent k-itemsets (S_k) for each k.

Suggested solution:

1. C_1: (A:6), (B:7), (C:7), (D:5), (E:6), (F:3).
   S_1: (A:6), (B:7), (C:7), (D:5), (E:6), (F:3).

2. C_2: (AB:4), (AC:4), (AD:2), (AE:4), (AF:1),
   (BC:4), (BD:4), (BE:4), (BF:2), (CD:3),
   (CE:5), (CF:3), (DE:1), (DF:2), (EF:1)
   S_2: (AB:4), (AC:4), (AD:2), (AE:4), (BC:4),
   (BD:4), (BE:4), (BF:2), (CD:3), (CE:5),
   (CF:3), (DF:2)

3. C_3: (ABC:2), (ABD:2), (ABE:2), (ACD:1),
   (ACE:3), (ADE, DE ~~not~~ frequent!), (BCD:2),
   (BCE:3), (BCF:2), (BDE, DE ~~not~~ frequent!)
   (BDF:1), (BEF, EF ~~not~~ frequent!),
   (CDE, DE ~~not~~ frequent!), (CDF:2),
   (CEF, EF ~~not~~ frequent!).
   S_3: (ABC:2), (ABD:2), (ABE:2), (ACE:3), (BCD:2), (BCE:3), (BCF:2), (CDF:2).

4. C_4: (ABCD, ACD ~~not~~ frequent!), (ABCE:1),
   (ABDE, ABE ~~not~~ frequent!), (BCDE, BDE ~~not~~ frequent!),
   (BCDF, BDF ~~not~~ frequent!), (BCEF, BEF ~~not~~ frequent!).
   S_4: $\emptyset$

Exercise 2-3 Color-histograms and distance functions (1 point)

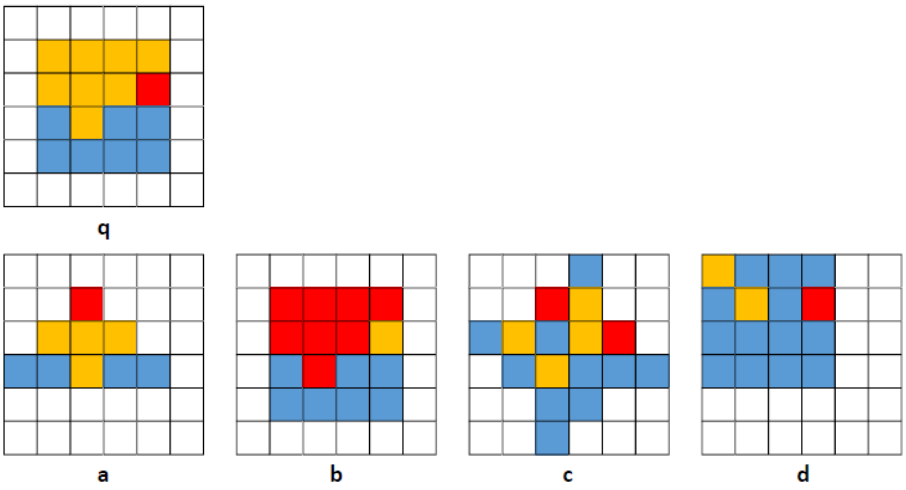
For each of the following distance measures (Euclidean, Manhattan, maximum, weighted Euclidean, quadratic form)

$$\begin{aligned}
\text{dist}_2(p,q) &= (|p_1-q_1|^2 + |p_2-q_2|^2 + |p_3-q_3|^2)^{\frac{1}{2}} \\
\text{dist}_1(p,q) &= |p_1-q_1| + |p_2-q_2| + |p_3-q_3| \\
\text{dist}_\infty(p,q) &= \max(|p_1-q_1|, |p_2-q_2|, |p_3-q_3|) \\
\text{dist}_w(p,q) &= (w_1|p_1-q_1|^2 + w_2|p_2-q_2|^2 + w_3|p_3-q_3|^2)^{\frac{1}{2}} \\
\text{dist}_M(p,q) &= ((p-q)M(p-q)^T)^{\frac{1}{2}}
\end{aligned}$$

calculate the distance between $p = (2,3,5)$ and $q = (4,7,8)$. As w use $(1, 1.5, 2.5)$ and as M use both of the following:

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad M_2 = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.8 \\ 0.7 & 0.8 & 1 \end{pmatrix}$$

Given 5 pictures with 36 pixels each.



Suggested solution:

$$\begin{aligned}
\text{dist}_2(p,q) &= 5.3851... \\
\text{dist}_1(p,q) &= 9 \\
\text{dist}_\infty(p,q) &= 4 \\
\text{dist}_w(p,q) &= 7.1063... \\
\text{dist}_{M1}(p,q) &= 5.3851... \\
\text{dist}_{M2}(p,q) &= 8.4261...
\end{aligned}$$

1. Extract from each picture a color histogram with the bins red, orange, and blue (the white pixels are ignored).

   Suggested solution:
   Color histograms (red, orange, blue); distance

$$
\begin{aligned}
q &= (1, 8, 7) \\
a &= (1, 4, 4) \\
b &= (8, 1, 7) \\
c &= (2, 4, 10) \\
d &= (1, 2, 13)
\end{aligned}
$$

2. Which pictures are most similar to the query q, using Euclidean distance? Give a ranking according to similarity to q.

   Suggested solution:
   Color histograms (red, orange, blue); distance

$$
\begin{aligned}
q = (1, 8, 7) \quad &; \\
a = (1, 4, 4) \quad &; \quad \text{dist}_2(q, a) = 5 \\
b = (8, 1, 7) \quad &; \quad \text{dist}_2(q, b) = 9.9 \\
c = (2, 4, 10) \quad &; \quad \text{dist}_2(q, c) = 5.1 \\
d = (1, 2, 13) \quad &; \quad \text{dist}_2(q, a) = 8.5
\end{aligned}
$$

   The ranking: a, c, d, b

3. The results are not entirely satisfactory. What could you change in the feature extraction or in the distance function to get better results? Report the improved feature extraction and features or the improved distance function.

   Suggested solution:
   Debatably, picture b is more similar to q than a or d are. The problem is that the Euclidean distance takes each color individually to compute the distance but does not take similarity between different colors (i.e., bins in the histogram) into account.

   A solution would be to use the quadratic form distance. We need a similarity matrix to define the (subjective) similarity of bins with each other:

$$
A = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}
$$

   $\text{dist}(q, a) = \sqrt{(q - a) \cdot A \cdot (q - a)^T} = 5$
   $\text{dist}(q, b) = 3.1$
   $\text{dist}(q, c) = 4.3$
   $\text{dist}(q, d) = 8.5$

Exercise 3-4 Visualization of distance functions (1 point)

Brainstorm on how you could visualize the effect of the different distance functions. To visualize the distance functions, try grouping a number of points as close or far w.r.t. a reference point. If you got some ideas: well, go and implement them! The point is credited for demonstrating a tool that you implemented yourself.

Suggested solution:

I would imagine a tool that produces a heatmap of distance values for the distances of points from the origin. You could treat the size of the two-dimensional plot $(x_{min}, x_{max}, y_{min}, y_{max})$, the resolution (granularity, step-size), and the distance function as parameters.