Solutions

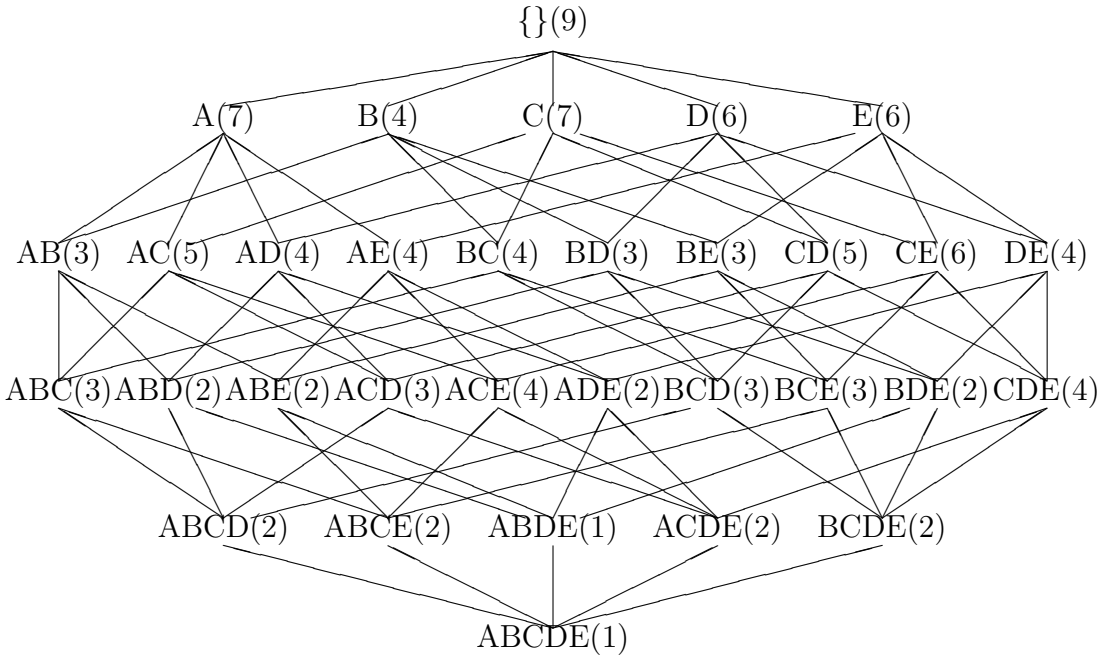
Exercise 3 : Closed Frequent Itemsets, Apriori, Color Histograms

Exercise 3-1 : Support based on closed frequent itemsets

(a) Draw a lattice diagram of the given database :

TID	A	B	C	D	E
1	1	0	0	1	0
2	1	1	1	0	1
3	1	0	1	0	1
4	1	1	1	1	1
5	0	0	1	1	1
6	1	0	1	1	1
7	1	0	0	0	0
8	0	1	1	1	1
9	1	1	1	1	0

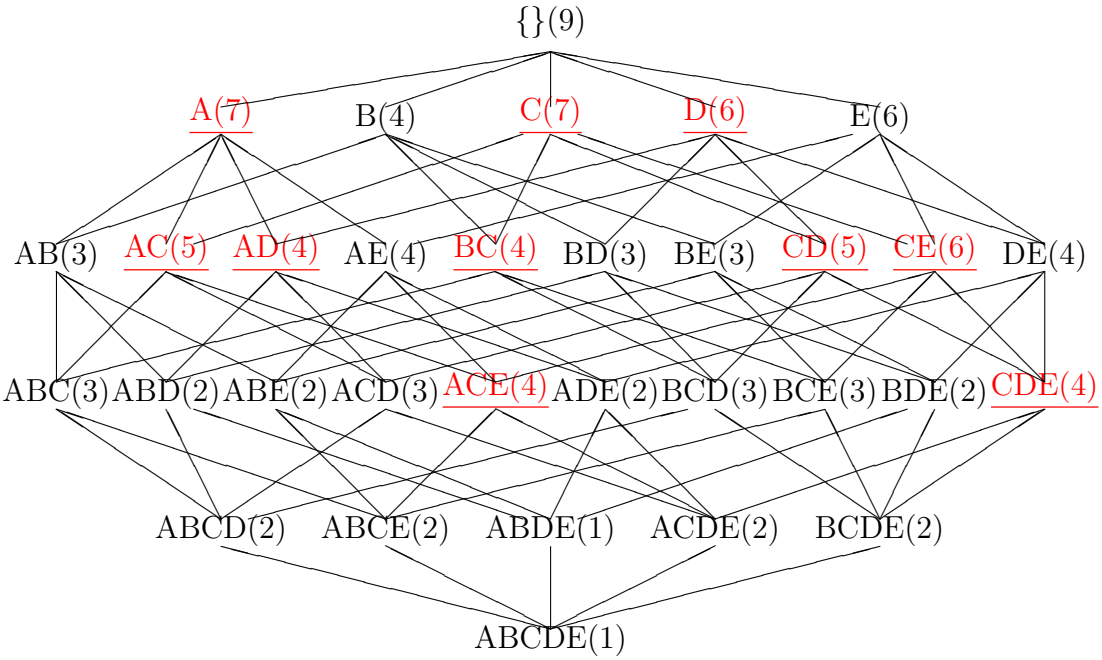
Suggested solution :



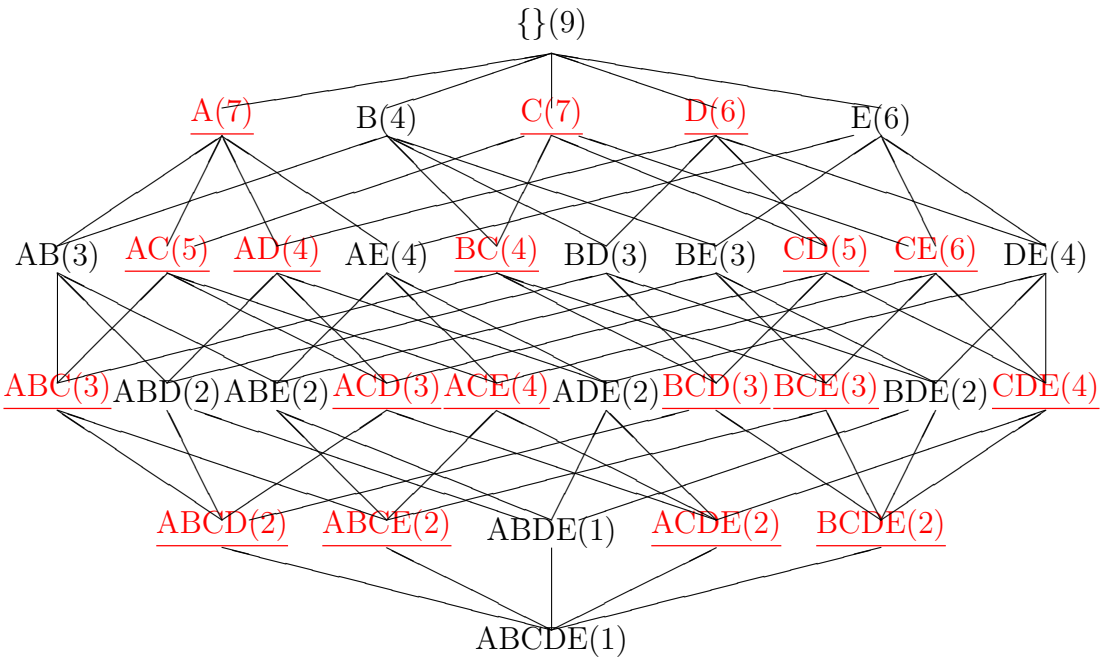
(b) Identify the closed frequent itemsets for the support thresholds $\sigma = 4$ and $\sigma = 2$, respectively. What do you observe ?

Suggested solution :

cfi for $\sigma = 4$:

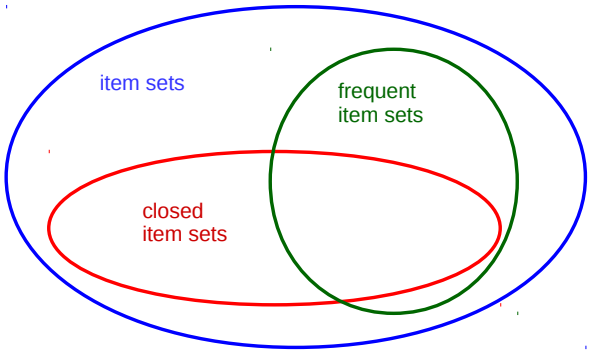


cfi for $\sigma = 2$:



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

Should be obvious considering the relation between itemsets, frequent itemsets, and closed itemsets :



(c) 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 :

C : set of closed frequent itemsets

$k_{\max}$: maximum size of closed frequent itemsets

$F_{\max} = \{f | f \in C, |f| = k_{\max}\}$ (find all frequent itemsets of size $k_{\max}$)

for $k = k_{\max} - 1$ to 1 do :

    $F_k = \{f | f \subset A, A \in F_{k+1}, |f| = k\}$ (find all frequent itemsets of size k)

    for each $f \in F_k$ do :

        if $f \notin C$ then

        $f_{\text{support}} = \max\{f'_{\text{support}} | f' \in F_{k+1}, f \subset f'\}$

        end if

    end for

end for

Exercise 3-2 : Apriori

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

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

Given the support threshold $\sigma = 2$, apply the Apriori algorithm and extract all frequent itemsets w.r.t. the given threshold. Please explain in the solution all the steps that you followed.

In particular include for each level the candidate set (C_k) (i) after the join step before pruning and (ii) after pruning. Annotate for those objects pruned in (ii) the explicit reason for pruning them.

Also give explicitly the solution of frequent k-itemsets (S_k) for each k.

Suggested solution :

C_1 : (A, 7) ; (B, 5) ; (C, 5) ; (D, 5) ; (E, 6) ; (F, 3) ; (G, 4)
S_1 : (A, 7) ; (B, 5) ; (C, 5) ; (D, 5) ; (E, 6) ; (F, 3) ; (G, 4)
C_2 :

(AB, 4)	(AC, 3)	(AD, 4)	(AE, 4)
(AF, 2)	(AG, 2)	(BC, 2)	(BD, 2)
(BE, 2)	~~(BF, 0)~~ (minSupport)	~~(BG, 1)~~ (minSupport)	~~(CD, 1)~~ (minSupport)
(CE, 4)	(CF, 2)	(CG, 2)	(DE, 3)
~~(DF, 1)~~ (minSupport)	~~(DG, 1)~~ (minSupport)	(EF, 3)	(EG, 2)
(FG, 2)			

S_2 : (AB, 4) ; (AC, 3) ; (AD, 4) ; (AE, 4) ; (AF, 2) ; (AG, 2) ; (BC, 2) ; (BD, 2) ; (BE, 2) ; (CE, 4) ; (CF, 2) ; (CG, 2) ; (DE, 3) ; (EF, 3) ; (EG, 2) ; (FG, 2)
C_3 :

(ABC, 2)	(ABD, 2)
(ABE, 2)	~~(ABF,)~~ (apriori property : BF not frequent !)
~~(ABG,)~~ (apriori property : BG not frequent !)	~~(ACD,)~~ (apriori property : CD not frequent !)
(ACE, 2)	~~(ACF, 1)~~ (minSupport)
~~(ACG, 1)~~ (minSupport)	(ADE, 2)
~~(ADF,)~~ (apriori property : DF not frequent !)	~~(ADG,)~~ (apriori property : DG not frequent !)
(AEF, 2)	~~(AEG, 1)~~ (minSupport)
~~(AFG, 1)~~ (minSupport)	~~(BCD,)~~ (apriori property : CD not frequent !)
~~(BCE, 1)~~ (minSupport)	~~(BDE, 1)~~ (minSupport)
(CEF, 2)	(CEG, 2)
(CFG, 2)	(EFG, 2)

S_3 : (ABC, 2) ; (ABD, 2) ; (ABE, 2) ; (ACE, 2) ; (ADE, 2) ; (AEF, 2) ; (CEF, 2) ; (CEG, 2) ; (CFG 2) ; (EFG, 2)
C_4 :
~~(ABCD,)~~ (apriori property : ACD, BCD not frequent !)
~~(ABCE,)~~ (apriori property : BCE not frequent !)
~~(ABDE,)~~ (apriori property : BDE not frequent !)
(CEFG, 2) S_4 : (CEFG, 2)
C_5 : $\emptyset$
S_5 : $\emptyset$

Exercise 3-3 : Color-histograms and distance functions

As a warm-up on distance measures : For each of the following distance measures (Euclidean, Manhattan, maximum, weighted Euclidean, quadratic form)

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

calculate the distance between $p=(2,4,7)$ and $q=(5,6,8)$. As w use $(2,2.5,3)$ and as M use both of the following :

$$M_1=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} \qquad M_2=\begin{pmatrix}1&0.8&0.6\\0.5&1&0.9\\0.5&0.7&1\end{pmatrix}$$

Suggested solution :

The correct solutions :

$$\begin{aligned}
\text{dist}_2(p,q) &= 3.7416\ldots \\
\text{dist}_1(p,q) &= 6 \\
\text{dist}_\infty(p,q) &= 3 \\
\text{dist}_w(p,q) &= 5.5677\ldots \\
\text{dist}_{M_1}(p,q) &= 3.7416\ldots \\
\text{dist}_{M_2}(p,q) &= 5.3197\ldots
\end{aligned}$$

Given 5 pictures as in Figure 1 with 36 pixels each.

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

(b) 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

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

Ranking : a, c, d, b

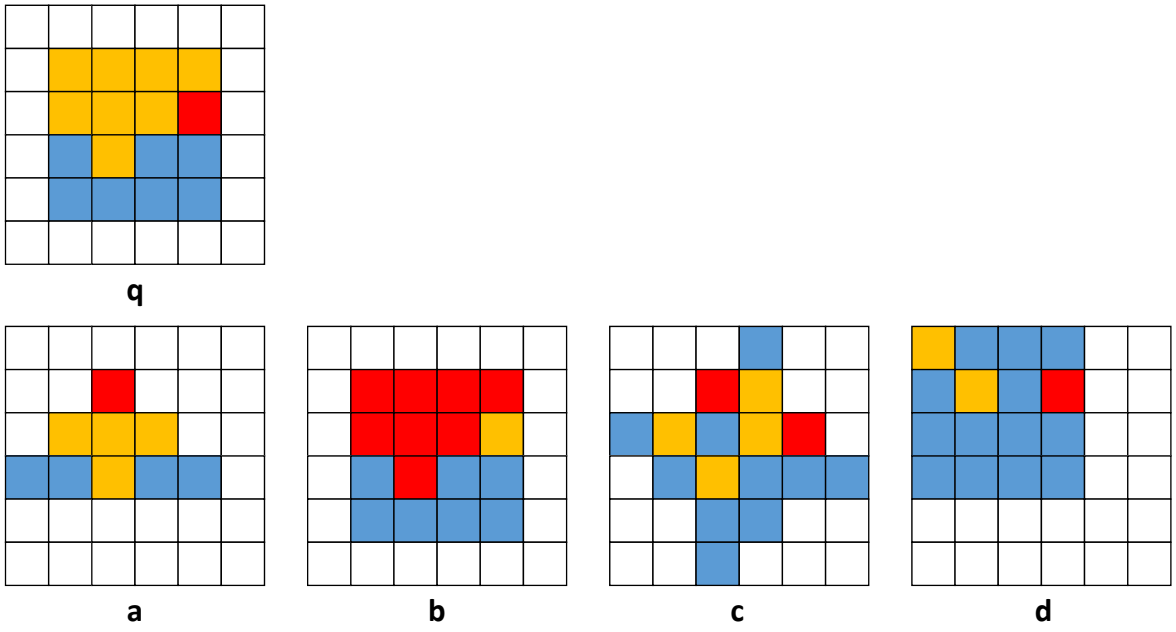


FIGURE 1 – 6×6 pixel pictures

(c) 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 Tool

Use T-distributed Stochastic Neighbor Embedding (T-SNE) tool for visualizing high-dimensional data which is a nonlinear dimensionality reduction technique to visualize data in a two or three dimensional space.

(a) Load python packages : `scikit-learn`, `seaborn`, `pandas`, `TSNE` from `sklearn.manifold`, and `load-digits` dataset from `sklearn.datasets`.

Suggested solution :

```
from sklearn.manifold import TSNE
from sklearn.datasets import load_digits
import seaborn as sns
import pandas as pd
```

(b) Load dataset and print dataset keys. Assign data as x and target as y, then investigate the shapes of the data.

Suggested solution :

```
mnist = load_digits()
print(mnist.keys())
x=mnist.data
y=mnist.target
print(x.shape)
print(y.shape)
```

(c) Define and fit the model by using the TSNE class.

Suggested solution :

```
tsne = TSNE(n_components=2, verbose=1, random_state=123)
z = tsne.fit_transform(x)
```

(d) Generate dataframe and plot the data.

Suggested solution :

```
#generate data frame
df = pd.DataFrame()
df["y"] = y
df["comp-1"] = z[:,0]
df["comp-2"] = z[:,1]
```

```
#plot data
sns.scatterplot(x="comp-1", y="comp-2", hue=df.y.tolist(),
                palette=sns.color_palette("hls", 10),
                data=df).set(title="MNIST data T-SNE projection")
plt.show()
```