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

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

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

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.

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)

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

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:

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

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

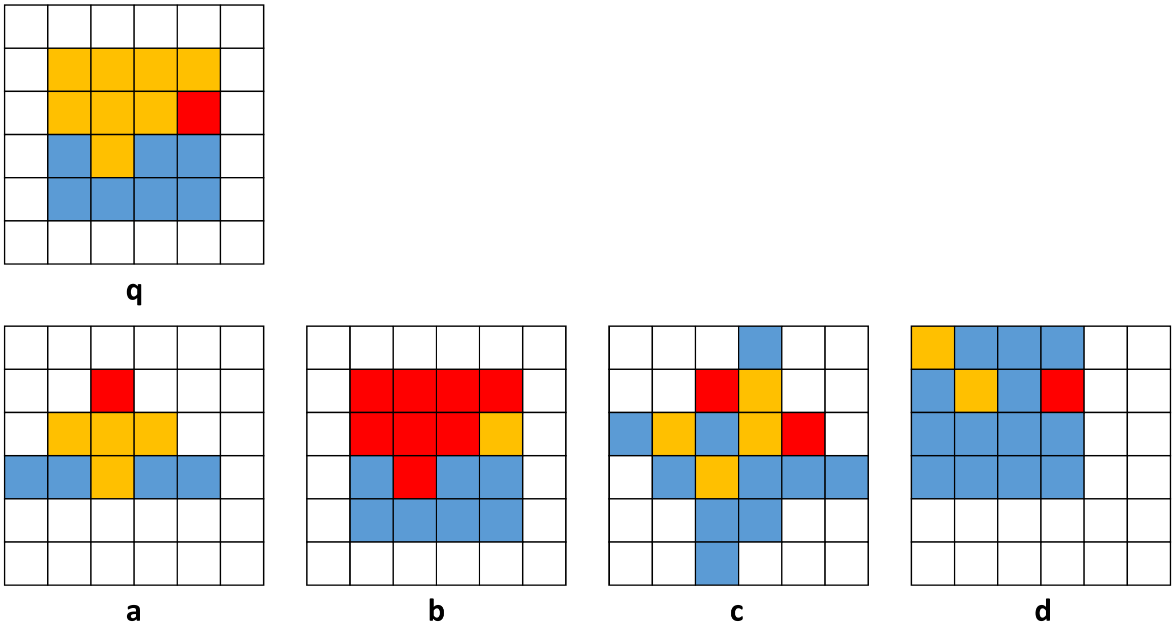

FIGURE 1 – 6 × 6 pixel pictures

(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.

(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.

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, numpy, seaborn, pandas`, `TSNE` from `sklearn.manifold`, and `load-digits` dataset from `sklearn.datasets`.

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

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

(d) Generate dataframe and plot the data.