CS-E4650 Methods of Data Mining

Jorma Laaksonen

EXAM 14.12.2022

These equipment are allowed and needed in the exam:

- pencil and eraser
- non-programmable calculator capable of roots, trigonometrics & logarithms

No other material (including phones, laptops, books, printouts or notes) is allowed.

Each task shall be written in separate sheets to ease and speed up the checking process!

1. Explain briefly, with 30–50 words, a mathematical definition and/or an illustration, the following concepts or abbreviations:

   a) similarity vs. distance vs. metric

   b) swap randomization

   c) closeness centrality

   d) PageRank

   e) tf-idf vector

2. Data items $t_1, \ldots, t_8$ have the following distance matrix:

	1	2	3	4	5	6	7	8
1	0	0.33	0.89	0.75	0.75	0.57	0.89	0.89
2	0.33	0	1.00	0.57	0.57	0.57	0.75	0.75
3	0.89	1.00	0	0.89	1.00	0.89	0.75	0.75
4	0.75	0.57	0.89	0	0.33	0.75	0.75	0.57
5	0.75	0.57	1.00	0.33	0	0.75	0.75	0.57
6	0.57	0.57	0.89	0.75	0.75	0	0.89	0.89
7	0.89	0.75	0.75	0.75	0.75	0.89	0	0.33
8	0.89	0.75	0.75	0.57	0.57	0.89	0.33	0

   a) Study clustering tendency based on the distance matrix. What threshold value would you choose for discriminating between intra- and inter-cluster distances?

   b) Apply agglomerative hierarchical clustering algorithm with single linkage metric on the data items. Draw an easily comprehensible dendrogram as the final result by re-arranging the data items appropriately.

   c) How do the shapes of the clusters generally differ between clusterting results with complete versus single linkage metric?

   d) How could you perform K-means clustering based on the same distance matrix? If not, explain why.

   e) Name and describe two clustering validation indices.

3. Consider the transaction database in the table below:

tid	items
1	a, c, d, e
2	a, d, e, f
3	b, c, d, e, f
4	b, d, e, f
5	b, e, f
6	c, d, e
7	c, e, f
8	d, e, f

a) Show the transactions in the form of a vertical "tid list" for all one-item sets.

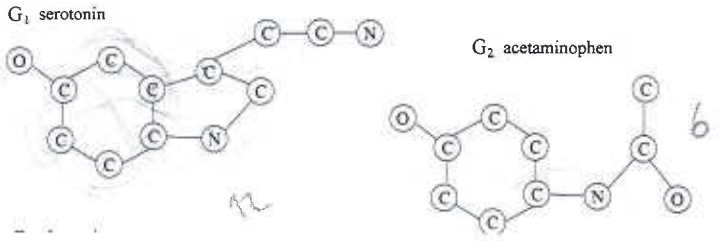
b) Apply the Apriori algorithm on the data with minimum frequency $\min_{\text{fr}} = 3/8$ and show all candidate and frequent itemsets in an enumeration tree.

c) Explain how one can prune itemset $\{c, d, f\}$ without frequency counting.

d) List all 1) maximal, 2) closed, and 3) 0-free frequent itemsets.

e) Calculate confidence (or precision) ϕ, leverage δ, and lift γ values of the following candidate rules: 1) $\{d\}\rightarrow\{f\}$, 2) $\{b\}\rightarrow\{f\}$, and 3) $\{b, e\}\rightarrow\{f\}$. Based on those values, explain which ones of the candidate rules have potential to express positive statistical dependence. How would you proceed in finding significant association rules?

4. Two chemical compounds are presented in the figure below:



a) Explain the concept and implementation of Maximum Common Subgraph (MCS) and apply it to $\mathbf{G}_1$ and $\mathbf{G}_2$.

b) Calculate value of max-normalized distance $Mdist(\mathbf{G}_1, \mathbf{G}_2) = 1 - \frac{|MCS(\mathbf{G}_1,\mathbf{G}_2)|}{\max\{|\mathbf{G}_1|,|\mathbf{G}_2|\}}$ and discuss its properties.

c) Explain the concept and implementation of Graph Edit Distance (GED) and apply it with equal unit costs to modify first $\mathbf{G}_1$ to $\mathbf{G}_2$ and then $\mathbf{G}_2$ to $\mathbf{G}_1$. Compare the results.

d) Use $\mathbf{G}_1$ and $\mathbf{G}_2$ as a database and identify and list all connected subgraphs of sizes $i = 1, 2, 3, 4$ and their frequencies.

e) How could you use frequent subgraphs for implementing a distance measure between two graphs?