# CS-E4650 Methods of Data Mining

Jorma Laaksonen

EXAM 22.2.2023

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.

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

   a) hierarchical clustering

   b) monotonicity of frequency

   c) the garden or forking paths

   d) the small-world problem

   e) bag-of-words model

2. A herd of cows have the individual attributes summarised in the table below.

| name | breed | age | milk production | character | music taste |
|---|---|---|---|---|---|
| Clover | Holstein | 2 | 20 | lively | rock |
| Sunny | Ayrshire | 2 | 10 | kind | rock |
| Rose | Holstein | 5 | 15 | calm | country |
| Daisy | Ayrshire | 4 | 25 | calm | classical |
| Strawberry | Finncattle | 7 | 35 | calm | classical |
| Molly | Ayrshire | 8 | 45 | kind | country |

   a) Which ones of the attributes are numerical and which are categorical? Further, which ones are nominal and which are ordinal?

   b) Explain K-means clustering and why it cannot be directly applied to data with non-numerical attributes.

   c) Explain K-modes clustering and how it can be applied to data with categorical attributes.

   d) Apply K-modes clustering to the above herd of cows by using only their categorical attributes and the value $K = 2$.

   e) Identify the stages of the K-modes algorithm that were dependent on the ordering of the data. What kinds of problems may follow from the dependency on data ordering?

3. Consider the candidate rules and their frequencies in the table below. The candidate rules are of the form $\mathbf{X} \rightarrow C{=}c$. $fr_X = fr(\mathbf{X})$, $fr_{XC} = fr(\mathbf{X}C{=}c)$.

| num | rule | $fr_X$ | $fr_{XC}$ | $fr_C$ | $\phi$ | $\delta$ | $\gamma$ | $nMI$ |
|---|---|---|---|---|---|---|---|---|
| 1 | smoking→CD | 250 | 100 | 300 | | 0.0250 | 1.33 | 11.07 |
| 2 | stress→CD | 250 | 75 | 300 | | 0.0000 | 1.00 | 0.00 |
| 3 | healthy diet→¬CD | 400 | 330 | 700 | | 0.0500 | 1.18 | 37.47 |
| 4 | regular doctor's visits→¬CD | 2 | 2 | 700 | 1.000 | | 1.43 | 1.03 |
| 5 | sun avoidance→¬CD | 300 | 215 | 700 | 0.717 | | 1.02 | 0.41 |
| 6 | female→¬CD | 500 | 360 | 700 | 0.720 | | 1.03 | 1.37 |
| 7 | smoking & sun avoidance→CD | 210 | 85 | 300 | 0.405 | 0.0220 | | 9.64 |
| 8 | no vaccine & no excercise→CD | 350 | 100 | 300 | 0.286 | -0.0050 | | 0.38 |
| 9 | smoking & healthy diet→CD | 210 | 80 | 300 | 0.381 | 0.0170 | | 5.80 |
| 10 | stress & smoking→CD | 90 | 60 | 300 | 0.667 | 0.0330 | 2.22 | |
| 11 | female & stress→CD | 130 | 60 | 300 | 0.462 | 0.0210 | 1.54 | |
| 12 | female & healthy diet→¬CD | 140 | 106 | 700 | 0.757 | 0.0080 | 1.08 | |

a) Show the equations for calculating the confidence (or precision) $\phi$, leverage $\delta$ and lift $\gamma$ values of the association rules.

b) Calculate the missing confidence, leverage, lift and $n$-normalized mutual information $nMI$ values in the table.

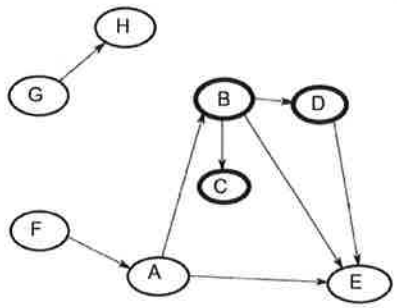
$$MI(\mathbf{X} \rightarrow C{=}c) = \log_2 \frac{P(\mathbf{X}C)^{P(\mathbf{X}C)} P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)} P(\neg\mathbf{X}C)^{P(\neg\mathbf{X}C)} P(\neg\mathbf{X}\neg C)^{P(\neg\mathbf{X}\neg C)}}{P(\mathbf{X})^{P(\mathbf{X})} P(\neg\mathbf{X})^{P(\neg\mathbf{X})} P(C)^{P(C)} P(\neg C)^{P(\neg C)}}$$

c) Which candidate rules would not be pruned out based on their leverage and lift values?

d) Which candidate rules would remain after further requiring $nMI \geq 1.5$?

e) What would be the next step in finding statistically significant association rules among those remaining after these steps?

4. The directed graph below depicts links between web pages named 'A',...,'H'. An arrow from e.g. node 'A' to 'B' means that the web page 'A' has a clickable link that points to 'B'.




a) Explain the "hubs and authorities" algorithm.

b) The root set $\mathbf{R}$ of a query consists of the bolded nodes 'B', 'C' and 'D'. Explain how this root set could have been obtained in a real scenario.

c) Form the base set $\mathbf{V}$ to be as large as possible. Which nodes belong to $\mathbf{V}$? What is the size $n$ of set $\mathbf{V}$? Initialize and show the hub weights $h_i$ and the authority weights $a_i$ of each page $i$ in $\mathbf{V}$.

d) Calculate the first iteration of the "hubs and authorities" algorithm.

e) Which nodes seem to be the best hubs and best authorities for the query?