Algorithmic Data Analysis

Esther Galbrun

Spring 2024

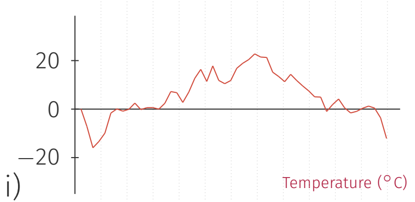
UNIVERSITY OF
EASTERN FINLAND

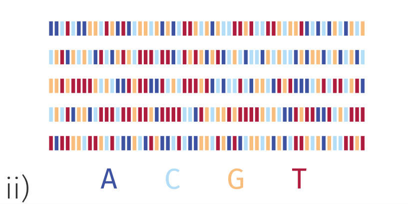
Last updated on 2023-12-28 11:37

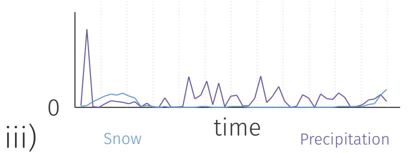
Associate characteristics to the datasets

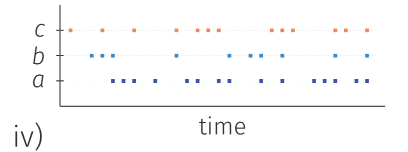
univariate vs. multivariate — sequential data vs. time-series

regularly vs. irregularly sampled — real vs. symbolic values









Q4.2: Distances (i)

We consider discrete sequences of items.

The distance between two sequence elements is defined as $d(s, s') = 0$ if and only if $s = s'$, 1 otherwise.

Given two sequences S and S', we denote

DTW(S, S') the *Dynamic Time Warping* (DTW) distance,

DTW$_w(S, S')$ the DTW distance with window constraint w, i.e. matching elements no further than w positions apart

E(S, S') the *Edit* distance with insertion and deletion operations such that $c_{\text{ins}} = 1$ and $c_{\text{del}} = 0.5$,

H(S, S') the *Hamming distance*,

L(S, S') the length of the *Longest contiguous common subsequence*.

Q4.2: Distances (i)

S_A and S_B are two sequences of length 10 such that $\mathrm{L}(S_A, S_B) = 5$ and S_C is a third sequence obtained by deleting the first and the last elements of S_A.

What can you say about the following statements?

i) $\mathrm{H}(S_A, S_B) \leq \mathrm{H}(S_C, S_B) + 2$
ii) $\mathrm{DTW}_3(S_A, S_B) < \mathrm{DTW}(S_A, S_B)$
iii) $\mathrm{DTW}(S_A, S_B) \leq \mathrm{H}(S_A, S_B)$
iv) $\mathrm{E}(S_A, S_B) \leq \mathrm{H}(S_A, S_B)$
v) $\mathrm{E}(S_A, S_B) < 8$
vi) $\mathrm{E}(S_A, S_C) \leq 1$
vii) E is a metric
viii) H is a metric
ix) L is a metric

Q4.3: Frequent sequences

We want to mine frequent sequences from a long sequence of itemsets, with minimum support threshold set to 4

The frequent sequences of length 3, with their support, are

$\{a, b, c\} : 7$	$\{a, b\}\{a\} : 4$	$\{a, b\}\{c\} : 5$	$\{a\}\{a, b\} : 6$
$\{a\}\{a, c\} : 8$	$\{a\}\{b, c\} : 6$	$\{b\}\{a, c\} : 5$	$\{b\}\{b, c\} : 4$

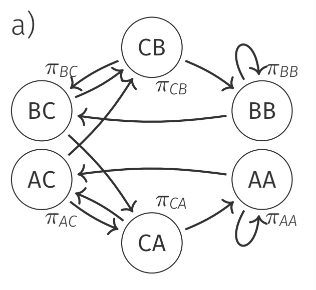
Provide the tightest possible upper bound on the support of each of the following sequences of length 4

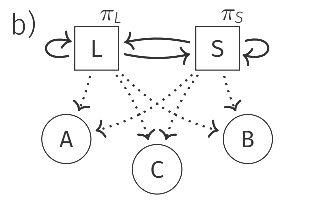
$$\text{supp}(\{a, b\}\{a, c\}) \leq ?$$
$$\text{supp}(\{a, b\}\{b, c\}) \leq ?$$
$$\text{supp}(\{a\}\{a, b, c\}) \leq ?$$
$$\text{supp}(\{a\}\{a, c\}\{a\}) \leq ?$$

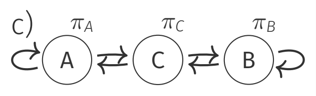
Q4.4: Markov models

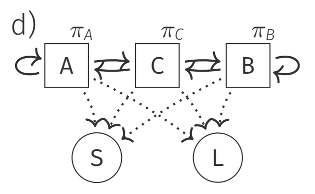
Associate each model to its name and to the size of the corresponding transition and/or emission matrices

Markov Chain   Hidden Markov Model   first/second order

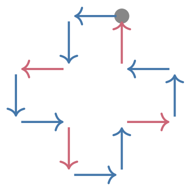




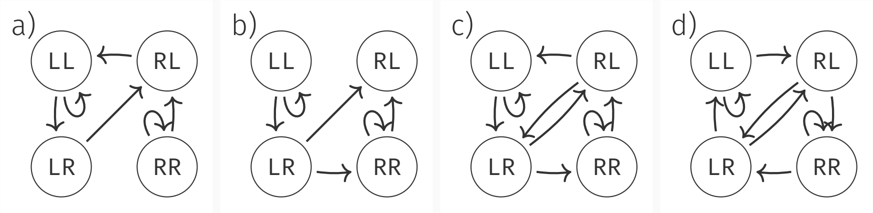




At each time step, $\mathcal{R}$ turns either left (L) or right (R) then rolls forward by one unit of distance. The path followed during one run, starting and ending at the gray the dot, facing up, is shown on the right.



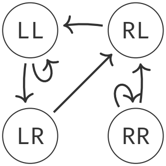
From which of the Markov chains below did it possibly arise?



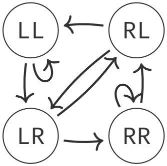
The path can be represented as sequence $S =$ LLRLLRLLRLLR

From which of the Markov chains below did it most likely arise?

a)



c)



$$\pi_a = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \quad (\text{LL, LR, RL, RR})$$

$$A_a = \begin{pmatrix} 0.90 & 0.10 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 \\ 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.90 & 0.10 \end{pmatrix}$$

$$\pi_c = \begin{pmatrix} 0.35 & 0.15 & 0.15 & 0.35 \end{pmatrix} \quad (\text{LL, LR, RL, RR})$$

$$A_c = \begin{pmatrix} 0.25 & 0.75 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.40 & 0.60 \\ 0.45 & 0.55 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.25 & 0.75 \end{pmatrix}$$

Q4.7: HMM problems

Match tasks, solutions and algorithms

Tasks

- Evaluation
- Explanation
- Training

Solutions

- $P_{\mathcal{M}}(O)$
- $\arg\max_{X \in \mathcal{X}} P_{\mathcal{M}}(X, O)$
- $\arg\max_{\mathcal{M} \in \mathcal{H}} P_{\mathcal{M}}(O)$

Algorithms

- Backward algorithm
- Baum–Welch algorithm
- Forward algorithm
- Viterbi algorithm