Algorithmic Data Analysis

Esther Galbrun
Autumn 2020
Q4.1: Data characteristics

Associate characteristics to the datasets

- univariate vs. multivariate
- regularly vs. irregularly sampled
- sequential data vs. time-series
- real vs. symbolic values

Temperature (°C)

a) Snow
b) Precipitation
c) A C G T
d)
Q4.2: Distances

We consider discrete sequences of items.

The distance between two sequence elements is defined as
\[ d(s, s') = 0 \text{ if and only if } s = s', \quad 1 \text{ otherwise.} \]

Given two sequences \( S \) and \( S' \), we denote

\[ \text{DTW}(S, S') \] the *Dynamic Time Warping* (DTW) distance,
\[ \text{DTW}_w(S, S') \] the DTW distance with window constraint \( w \),
\[ \text{E}(S, S') \] the *Edit* distance with insertion and deletion operations such that \( c_{\text{ins}} = 1 \) and \( c_{\text{del}} = 0.5 \),
\[ \text{H}(S, S') \] the *Hamming distance*,
\[ \text{L}(S, S') \] the length of the *Longest contiguous common subsequence*. 


Q4.2: Distances

$S_A$ and $S_B$ are two sequences of length 10 such that $L(S_A, S_B) = 5$ 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?

a) $H(S_A, S_B) \leq H(S_C, S_B) + 2$

b) $DTW_3(S_A, S_B) < DTW(S_A, S_B)$

c) $DTW(S_A, S_B) \leq H(S_A, S_B)$

d) $E(S_A, S_B) \leq H(S_A, S_B)$

e) $E(S_A, S_B) < 8$

f) $E(S_A, S_C) \leq 1$

g) $E$ is a metric

h) $H$ is a metric

i) $L$ 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

a) 

b) 

c) 

d)
Q4.5: Markov unchained

At each time step, 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 dot, facing up, is shown on the right.

From which of the Markov chains below did it possibly arise?
Q4.6: Markov unchained (continued)

The path can be represented as sequence $S = \text{LLRLLRLLRLLR}$

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

![Markov Chains](image)

- a) $\pi_a = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.90 & 0.10 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 \\ 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.90 & 0.10 \end{pmatrix}$

- $A_a = \begin{pmatrix} 0.25 & 0.75 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.40 & 0.60 \\ 0.45 & 0.55 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.25 & 0.75 \end{pmatrix}$
Match tasks, solutions and algorithms

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation</td>
<td>$P_M(O)$</td>
</tr>
<tr>
<td>Explanation</td>
<td>$\arg\max_{X \in \mathcal{X}} P_M(X, O)$</td>
</tr>
<tr>
<td>Training</td>
<td>$\arg\max_{M \in \mathcal{H}} P_M(O)$</td>
</tr>
</tbody>
</table>

Algorithms

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