Task #1: POS-tagging and Viterbi decoding

Given the sentence "fish swim" and the following probabilities:

tags: N (noun) and V (verb) with initial probabilities: P(N)=0.6, P(V)=0.4

transition:

- P(N → N) = 0.2   P(N → V) = 0.8
- P(V → N) = 0.5   P(V → V) = 0.5

emission:

- P(fish|N) = 0.7   P(fish|V) = 0.1
- P(swim|N) = 0.1   P(swim|V) = 0.4
- P(another word | N)   P(another word | V)
- …

Run one full iteration of the Viterbi algorithm and produce:

- the Viterbi decoding matrix C (values for each word-tag pair)
- the final most-likely tag sequence

Viterbi Algorithm

for $l \in Q$

$V_1(l) = B_{l,x_1} \cdot S_l$

for i=1,…,n-1

for $l \in Q$

$V_{i+1}(l) = B_{l,x_{i+1}} \cdot \max_{k \in Q}\{V_i(k) \cdot A_{k,l}\}$

$P_{i+1}(l) = \arg\max_{k \in Q}\{V_i(k) \cdot A_{k,l}\}$

$h_n^* = \arg\max_{l \in Q}\{V_n(l)\}$

for i=n-1,…,1

$h_i^* = P_{i+1}(h_{i+1}^*)$

הגדרה

Hidden Markov Model מורכב מהשלישיה (Σ, Q, θ) כאשר:

- Σ - אלף בית של סמלים.
- Q - קבוצה סופית של מצבים, בכל מצב ייתכנו מספר סמלי פלט.
- θ - קבוצה של הסתברויות:
  - הסתברויות מעבר בין המצבים $A_{k,l} = P(H_{i+1} = l \mid H_i = k)$
  - הסתברויות פלט (הפלט תלוי אך ורק במצב שבו אנו נמצאים) $B_{k,a} = P(X_i = a \mid H_i = k)$

Transition matrix

A

$$A = \begin{array}{c|cc} & N & V \\ \hline N & 0.2 & 0.8 \\ V & 0.5 & 0.5 \end{array}$$

Initial distribution vector

$$\begin{array}{cc} N & V \end{array}$$

$$\pi = (0.6, 0.4)$$

Emission matrix:

B

$$B = \begin{array}{c|cc} & \text{fish} & \text{swim} \\ \hline N & 0.7 & 0.1 \\ V & 0.1 & 0.4 \end{array}$$

$V_1(N) = B_{N\,fish} \cdot S_N = 0.7 \cdot 0.6 = 0.42$

$V_1(V) = B_{V\,fish} \cdot S_V = 0.1 \cdot 0.4 = 0.04$

$V_2(N) = B_{N\,swim} \cdot \max\{V_1(N) \cdot A_{N,N}, V_1(V) \cdot A_{V,N}\} =$

$0.1 \cdot \max\{0.42 \cdot 0.2, 0.04 \cdot 0.5\} =$

$0.1 \cdot \max\{0.084, 0.02\} = 0.1 \cdot 0.082 = 0.0084$

$P_2(N) = N$

$V_2(V) = B_{V\,swim} \cdot \max\{V_1(N) \cdot A_{NV}, V_1(V) \cdot A_{VV}\} =$

$0.4 \cdot \max\{0.42 \cdot 0.8, 0.04 \cdot 0.5\} =$

$0.4 \cdot \max\{0.336, 0.02\} = 0.4 \cdot 0.336 = 0.1344$

$P_2(V) = N$

C

V_{mat}	1	2
N	0.42	0.0084
V	0.04	0.1344

P_{mat}	1	2
N		N
V		N

$h_2 = \arg\max\{V_2(N), V_2(V)\} =$

$\arg\max\{\overset{N}{0.0082}, \overset{V}{0.1344}\} = V$

$h_1 = P_2(h_2) = P_2(V) = N$

The most likely hidden state sequence = N, V