Supervision 3

Markov assumption

Transitions depend only on the current state. The probability of an output observation depends only on the current state.

The assumptions are practically important because it allows us to model the transition/output probabilities as matrices rather than higher-dimensional tensors, since they are memoryless.

If this were not the case we would not be able to ascribe a value to a_{ij} because you would need to know the state history, and likewise for $b_i(k_j)$

HMM Artificial data

1

A list of states, a list of outputs, a transition probability matrix and an output probability matrix.

2

1. Start in state `Start`.

2. Output observation `Start`.

3. Sample the probability distribution of possible states to which to transition from the current state. Update the current state accordingly

4. If you are currently in state `End`, output observation `End`, then terminate.

5. Otherwise, sample the probability distribution of possible output observations from the current state. Output accordingly

6. Repeat from step 3.

3

It could be the case that you are more likely to transition to the end state from some states than you are from others. For example imagine an HMM modelling a shop with states `Start`, `Shop open`, `Shop closed`, and `End`. If the HMM is supposed to model an average day, it is far more likely that the day would end while the shop is closed as opposed to when it's open.

Smoothing in HMMs

1

It is counterproductive to smooth when certain transitions/outputs are impossible from some state. This is because smoothing would ascribe these transitions/outputs a non-zero probability. Smoothing is only applicable when just because a transition does not happen in the training data, you do not want to make the assumption that it can never happen. Likewise for output observations.

2

The observations are a better candidate for smoothing. This is because the likelihood of an amino acid which we have previously never seen, for example, inside the membrane to be observed there is greater than the probability of, for example, the protein jumping straight from inside the cell to outside the cell without going through the membrane

Viterbi and Forward algorithm

1

$\alpha_t(j) = \Sigma_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t)$

The forward trellis is the probability of observing a given set of observations up to time t and also the current state at time t being state j.

This is equal to the probability of seeing all of the observations up to time t-1 from the previous state, times the probability of emitting the t^{th} observation from state j, times the probability of transitioning to state j given the previous state. However, since we don't know what the previous state is, we have to iterate over all possible previous states i, and multiply each term by the probability of the previous state being i.

Therefore the forward trellis at time t of state j is equal to the sum over all possible previous states i of the probability of transitioning from i to j, times the probability of emitting the t^{th} observation from state j, times the probability of the previous state being i, times the probability of seeing all of the observations up to time t-1. The last two of these terms can be grouped into the forward trellis at time t-1 of state i.

This is the same form as the recursive formula above.

2

It is not certain which state was the previous one, and so you have to do a weighted sum over all of the possibilities, weighted according to the probability of that path.

Parts of Speech tagging with HMM

1

We (personal pronoun) can (auxiliary verb) fish (verb) (i.e. we are able to fish)

We (personal pronoun) can (verb) fish (noun) (i.e. we put fish in a can)

2

$P(X_a) = P(s_0) \times P(s_3|s_0) \times P(s_4|s_3) \times P(s_1|s_4) \times P(s_f|s_1)$

$= a_{03} \times a_{34} \times a_{41} \times a_{1f}$

$= 0.60 \times 0.40 \times 0.73 \times 0.15 \approx 0.0263$

$P(X_b) = P(s_0) \times P(s_3|s_0) \times P(s_1|s_3) \times P(s_2|s_1) \times P(s_f|s_2)$

$= a_{03} \times a_{31} \times a_{12} \times a_{2f}$

$= 0.60 \times 0.40 \times 0.63 \times 0.20$

≈ 0.0302

$P(X_a,O) = P(X_a) \times P(O|X_a)$

$= P(X_a) \times P(we|s_3) \times P(can|s_4) \times P(fish|s_1)$

$= a_{03} \times a_{34} \times a_{41} \times a_{1f} \times b_3(we) \times b_4(can) \times b_1(fish)$

$= 0.60 \times 0.40 \times 0.73 \times 0.15 \times 1 \times 1 \times 0.89$

≈ 0.0234

$P(X_b,O) = P(X_b) \times P(O|X_b)$

$= P(X_b) \times P(we|s_3) \times P(can|s_1) \times P(fish|s_2)$

$= a_{03} \times a_{31} \times a_{12} \times a_{2f} \times b_3(we) \times b_1(can) \times b_2(fish)$

$= 0.60 \times 0.40 \times 0.63 \times 0.20 \times 1 \times 0.10 \times 0.75$

≈ 0.00227

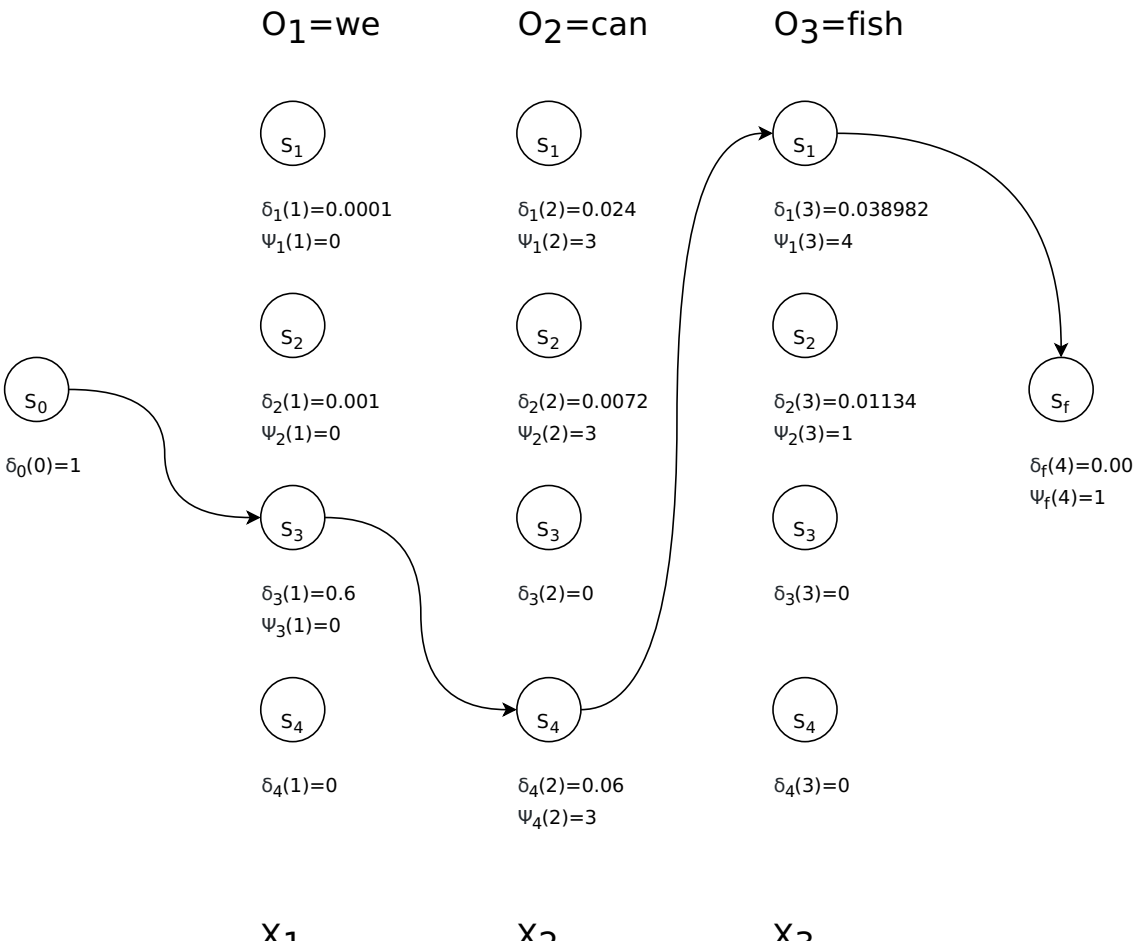
$P(X_a)$ and $P(X_b)$ are used in the HMM

3

Here $\delta_i(t)$ is the probability of the most likely path from $0 \le$ time $\le t$ for which the state at time t is s_i.

$\Psi_i(t)$ is the most likely predecessor of state s_i at time t.

These are the same definitions as in the lecture.



For every possible second state, it is most likely that the first observation (we) was a personal pronoun.

However, if the last observation (fish) was a verb, it is most likely that the previous observation (can) is a verb. Whereas if "fish" was a noun, it is most likely that "can" is an auxiliary verb. These are the only two possible states for "fish".

4

The estimated sequence is s_0,s_3,s_4,s_1,s_f (i.e. we are able to fish) which I would say is the correct interpretation.

5

$b_i(z)$ = proportion of the data labelled with state s_i which is also labelled with observation z.

6

Some states overlap when it comes to words. For example any word which is a verb is also an auxiliary verb. Likewise any proper noun is also a noun.

The problem could be fixed by making sure that the states are mutually exclusive (e.g. verb → non-auxiliary verb, noun →non-proper noun in the cases above)

Viterbi with higher order HMMs

1

N^2

2

For a k-order HMM, the time complexity is O(kT). This is linear in the case of constant k but if k=N then the time complexity is O(NT).