HMMs

for sequence labelling

Thomas Brochhagen // Session 3, 2023 // NLP@UPF

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state q_i
$\pi = \pi_1, \pi_2, ..., \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

Markov assumption $$P(q_i \mid q_1, \ldots, q_{i-1}) = P(q_i \mid q_{i-1})$$

Output Independence $$P(o_i \mid q_1, \ldots, q_T; o_1, \ldots, o_T) = P(o_i \mid q_i)$$

A-matrix
(tag transitions)

$$P(t_i \mid t_{i-1}) = \frac{\text{Count}(t_{i-1}, t_i)}{\text{Count}(t_{i-1})}$$

A-matrix
(tag transitions)

$$P(t_i \mid t_{i-1}) = \frac{\text{Count}(t_{i-1}, t_i)}{\text{Count}(t_{i-1})}$$

B-matrix
(tag to word)

$$P(w_i \mid t_i) = \frac{\text{Count}(t_i, w_i)}{\text{Count}(t_i)}$$

Decoding: Given as input an HMM $\lambda = (A, B)$ and a sequence of observations $O = o_1, o_2, ..., o_T$, find the most probable sequence of states $Q = q_1 q_2 q_3 \ldots q_T$.

Most probable tag-sequence given word-sequence

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(t_1, \ldots, t_n \mid w_1, \ldots w_n)$$

Most probable tag-sequence given word-sequence

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(t_1, \ldots, t_n \mid w_1, \ldots w_n)$$

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} \frac{P(w_1, \ldots w_n \mid t_1, \ldots t_n) P(t_1, \ldots, t_n)}{P(w_1, \ldots, w_n)}$$

Most probable tag-sequence given word-sequence

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(t_1, \ldots, t_n \mid w_1, \ldots w_n)$$

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} \frac{P(w_1, \ldots w_n \mid t_1, \ldots t_n) P(t_1, \ldots, t_n)}{P(w_1, \ldots, w_n)}$$

By assumption $$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(w_1, \ldots w_n \mid t_1, \ldots t_n) P(t_1, \ldots, t_n)$$

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(w_1, \ldots w_n \mid t_1, \ldots t_n) P(t_1, \ldots, t_n)$$

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(w_1, \ldots w_n \mid t_1, \ldots t_n) P(t_1, \ldots, t_n)$$

By assumption

$$P(w_1, \ldots, w_n \mid t_1, \ldots, t_n) = \Pi_i P(w_i \mid t_i)$$

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(w_1, \ldots w_n \mid t_1, \ldots t_n) P(t_1, \ldots, t_n)$$

By assumption $$P(w_1, \ldots, w_n \mid t_1, \ldots, t_n) = \Pi_i P(w_i \mid t_i)$$

By assumption $$P(t_1, \ldots, t_n) \approx \Pi_i P(t_i \mid t_{i-1})$$

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} P(w_1, \ldots w_n \mid t_1, \ldots t_n) P(t_1, \ldots, t_n)$$

By assumption $$P(w_1, \ldots, w_n \mid t_1, \ldots, t_n) = \Pi_i P(w_i \mid t_i)$$

By assumption $$P(t_1, \ldots, t_n) \approx \Pi_i P(t_i \mid t_{i-1})$$

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} \Pi_i \, P(w_i \mid t_i) P(t_i \mid t_{i-1})$$

Transition (A-matrix)

$$\hat{t}_{1:n} = \text{argmax}_{t_1,\ldots,t_n} \Pi_i \, P(w_i \mid t_i) \;\; P(t_i \mid t_{i-1})$$

Emission (B-matrix)