Linguistics 696f
Hammond
Spring '03

Algorithms for Hidden Markov Models

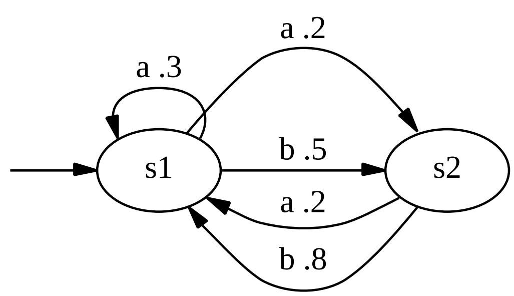
A. Overview

(1)
a. Viterbi algorithm
b. Forward probabilities
c. Backward probabilities
d. Training a simple Markov chain
e. Baum-Welch

B. Viterbi

(2) $\arg\max_{s_{1,t}} \mathrm{p}(s_{1,t}|w_{1,t-1})$

(3) Multiple arcs



(4)

	[]
s_1	$\emptyset, 1$
s_2	$\emptyset, 0$

(5)

	[]	$[a]$
s_1	$\emptyset, 1$	$s_1, .3$
s_2	$\emptyset, 0$	$s_1, .2$

(6)

	[]	$[a]$	$[a\ a]$
s_1	$\emptyset, 1$	$s_1, .3$	$s_1 \rightarrow s_1, .04$
s_2	$\emptyset, 0$	$s_1, .2$	$s_1 \rightarrow s_1, .01$

(7)

	[]	[a]	[a a]	[a a a]
s_1	$\emptyset, 1$	$s_1, .3$	$s_1 \rightarrow s_1, .04$	$s_1 \rightarrow s_1 \rightarrow s_1, .012$
s_2	$\emptyset, 0$	$s_1, .2$	$s_1 \rightarrow s_1, .01$	$s_1 \rightarrow s_1 \rightarrow s_1, .008$

(8) $\sigma_i(t)$, the most likely state sequence that generates $w_1 \ldots w_{t-1}$ ending up in state s_i.

(9) $\sigma_i(t) \stackrel{\text{def}}{=} \arg\max_{s_{1,t}} \mathrm{p}(w_{1,t-1}, s_{1,t-1}, S_t = s^i)$

(10) $\sigma_i(1) = s^i$ if $t = 1$

(11) $\sigma_i(t+1) = \sigma_j(t) \circ s^i,$ where $j = \arg\max_{k=1}^{\sigma} \mathrm{p}(\sigma_k(t))\mathrm{p}(s^k \stackrel{w_t}{\rightarrow} s^i)$ if $t > 1$

C. Forward probabilities

(12) $\alpha_i(t)$ the probability of producing $w_{1,t-1}$ and ending up in s^i.

(13) $\alpha_i(t) \stackrel{\text{def}}{=} \mathrm{p}(w_{1,t-1}, S_t = s^i),$ where $t > 1$

(14) $\alpha_i(1) = \pi_i$

(15) $\alpha_j(t+1) = \sum_{i=1}^{\sigma} \alpha_i(t)\mathrm{p}(s^i \stackrel{w_t}{\rightarrow} s^j)$

(16) What is the forward probability of [*a b a*] given the HMM above?

(17)

	[]
s_1	1
s_2	0

(18)

$$\begin{aligned} \alpha_1(2) &= \alpha_1(1)\mathrm{p}(s^1 \stackrel{a}{\rightarrow} s^1) + \alpha_2(1)\mathrm{p}(s^2 \stackrel{a}{\rightarrow} s^1) \\ &= (1 \times .3) + (0 \times .2) \\ &= .3 \end{aligned}$$

(19)

$$\begin{aligned} \alpha_2(2) &= \alpha_1(1)\mathrm{p}(s^1 \stackrel{a}{\rightarrow} s^2) + \alpha_2(1)\mathrm{p}(s^2 \stackrel{a}{\rightarrow} s^2) \\ &= (1 \times .2) + (0 \times 0) \\ &= .2 \end{aligned}$$

(20)

	[]	[a]
s_1	1	.3
s_2	0	.2

(21)

$$\begin{aligned} \alpha_1(3) &= \alpha_1(2)\mathrm{p}(s^1 \stackrel{b}{\rightarrow} s^1) + \alpha_2(2)\mathrm{p}(s^2 \stackrel{b}{\rightarrow} s^1) \\ &= (.3 \times 0) + (.2 \times .8) \\ &= .16 \end{aligned}$$

(22)
$$\begin{aligned} \alpha_2(3) &= \alpha_1(2)\mathrm{p}(s^1 \xrightarrow{b} s^2) + \alpha_2(2)\mathrm{p}(s^2 \xrightarrow{b} s^2) \\ &= (.3 \times .5) + (.2 \times 0) \\ &= .15 \end{aligned}$$

(23)

	[]	[a]	[a b]
s_1	1	.3	.16
s_2	0	.2	.15

(24)
$$\begin{aligned} \alpha_1(4) &= \alpha_1(3)\mathrm{p}(s^1 \xrightarrow{a} s^1) + \alpha_2(3)\mathrm{p}(s^2 \xrightarrow{a} s^1) \\ &= (.16 \times .3) + (.15 \times .2) \\ &= .078 \end{aligned}$$

(25)
$$\begin{aligned} \alpha_2(4) &= \alpha_1(3)\mathrm{p}(s^1 \xrightarrow{a} s^2) + \alpha_2(3)\mathrm{p}(s^2 \xrightarrow{a} s^2) \\ &= (.16 \times .2) + (.15 \times 0) \\ &= .032 \end{aligned}$$

(26)

	[]	[a]	[a b]	[a b a]
s_1	1	.3	.16	.078
s_2	0	.2	.15	.032

D. Backward probabilities

(27) $\beta_i(t)$ refers to the overall probability of producing $w_{t,n}$ where the HMM is in state s^i at time t.

(28) $\beta_i(n+1) = \mathrm{p}(\epsilon | S_{n+1} = s^i) = 1$

(29) $\beta_i(t-1) = \sum_{j=1}^{\sigma} \mathrm{p}(s^j \xrightarrow{w_{t-1}} s^j)\beta_j(t)$

(30) How do we calculate the *backward* probability of the same string [a b a] with the same HMM?

(31)

	[]
s_1	1
s_2	1

(32)
$$\begin{aligned} \beta_1(3) &= \mathrm{p}(s^1 \xrightarrow{a} s^1)\beta_1(4) + \mathrm{p}(s^1 \xrightarrow{a} s^2)\beta_2(4) \\ &= (.3 \times 1) + (.2 \times 1) \\ &= .5 \end{aligned}$$

(33)
$$\begin{aligned} \beta_2(3) &= \mathrm{p}(s^2 \xrightarrow{a} s^1)\beta_1(4) + \mathrm{p}(s^2 \xrightarrow{a} s^2)\beta_2(4) \\ &= (.2 \times 1) + (0 \times 1) \\ &= .2 \end{aligned}$$

(34)

	[]	[a]
s_1	1	.5
s_2	1	.2

(35)
$$\begin{aligned} \beta_1(2) &= \mathrm{p}(s^1 \xrightarrow{b} s^1)\beta_1(3) + \mathrm{p}(s^1 \xrightarrow{b} s^2)\beta_2(3) \\ &= (0 \times .5) + (.5 \times .2) \\ &= .1 \end{aligned}$$

(36)
$$\begin{aligned} \beta_2(2) &= \mathrm{p}(s^2 \xrightarrow{b} s^1)\beta_1(3) + \mathrm{p}(s^2 \xrightarrow{b} s^2)\beta_2(3) \\ &= (.8 \times .5) + (0 \times .2) \\ &= .4 \end{aligned}$$

(37)

	[]	[a]	[b a]
s_1	1	.5	.1
s_2	1	.2	.4

(38)
$$\begin{aligned} \beta_1(1) &= \pi_1\mathrm{p}(s^1 \xrightarrow{a} s^1)\beta_1(2) + \pi_1\mathrm{p}(s^1 \xrightarrow{a} s^2)\beta_2(2) \\ &= (1 \times .3 \times .1) + (1 \times .2 \times .4) \\ &= .11 \end{aligned}$$

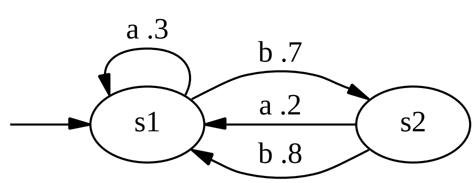
(39)
$$\begin{aligned} \beta_2(1) &= \pi_2\mathrm{p}(s^2 \xrightarrow{a} s^1)\beta_1(2) + \pi_2\mathrm{p}(s^2 \xrightarrow{a} s^2)\beta_2(2) \\ &= (0 \times .2 \times .1) + (0 \times 0 \times .4) \\ &= 0 \end{aligned}$$

(40)

	[]	[a]	[b a]	[a b a]
s_1	1	.5	.1	.11
s_2	1	.2	.4	0

E. Training a simple Markov chain

(41) A simple Markov chain



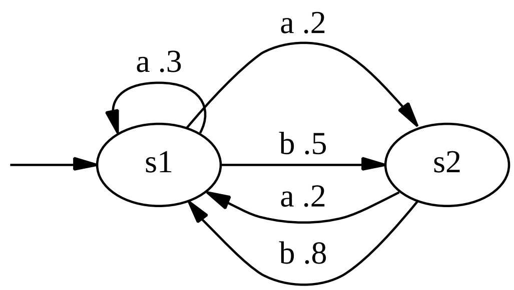
(42) $$\mathrm{p}_e(s^i \xrightarrow{w^k} s^j) = \frac{|s^i \xrightarrow{w^k} s^j|}{\sum_{l=1,m=1}^{\sigma,\omega} |s^i \xrightarrow{w^m} s^l|}$$

(43) $$s^1 \xrightarrow{a} s^1 \xrightarrow{b} s^2 \xrightarrow{a} s^1 \xrightarrow{b} s^2 \xrightarrow{a} s^1 \xrightarrow{b} s^2 \xrightarrow{b} s^1$$

	arc	count	probability
	$s^1 \xrightarrow{a} s^1$	1	.25
(44)	$s^1 \xrightarrow{b} s^2$	3	.75
	$s^2 \xrightarrow{a} s^1$	2	.66
	$s^2 \xrightarrow{b} s^1$	1	.33

F. Baum-Welch

(45) Multiple arcs



(46) $|s^i \xrightarrow{w^k} s^j| = \frac{1}{\mathrm{p}(w_{1,n})} \sum_{t=1}^{n} \alpha_i(t) \mathrm{p}(s^i \xrightarrow{w^k} s^j) \beta_j(t+1)$

(47) We will train the HMM above with *aabb*. We first calculate the values for α and β.

(48)

	$\alpha(1)$	$\alpha(2)$	$\alpha(3)$	$\alpha(4)$	$\alpha(5)$
s^1	1	.3	.13	.048	.052
s^2	0	.2	.06	.065	.024

(49)

	$\beta(1)$	$\beta(2)$	$\beta(3)$	$\beta(4)$	$\beta(5)$
s^1	.076	.2	.4	.5	1
s^2	0	.08	.4	.8	1

(50)

$\|s^i \xrightarrow{w^k} s^j\|$	=	$\frac{1}{\mathrm{p}(w_{1,n})} \sum_{t=1}^{n} \alpha_i(t) \mathrm{p}(s^i \xrightarrow{w^k} s^j) \beta_j(t+1)$		
$\|s^1 \xrightarrow{a} s^1\|$	=	$\frac{1}{.074}[(1 \times .3 \times .2) + (.3 \times .3 \times .4)]$	=	1.2972
$\|s^1 \xrightarrow{a} s^2\|$	=	$\frac{1}{.074}[(1 \times .2 \times .08) + (.3 \times .2 \times .4)]$	=	.5405
$\|s^1 \xrightarrow{b} s^2\|$	=	$\frac{1}{.074}[(.13 \times .5 \times .8) + (.048 \times .5 \times 1)]$	=	1.027
$\|s^2 \xrightarrow{a} s^1\|$	=	$\frac{1}{.074}[(0 \times .2 \times .2) + (.2 \times .2 \times .4)]$	=	.2162
$\|s^2 \xrightarrow{b} s^1\|$	=	$\frac{1}{.074}[(.06 \times .8 \times .5) + (.065 \times .8 \times 1)]$	=	1.027

(51)

$\lvert s^i \xrightarrow{w^k} s^j \rvert$	p_e
$\lvert s^1 \xrightarrow{a} s^1 \rvert$	.45
$\lvert s^1 \xrightarrow{a} s^2 \rvert$	.19
$\lvert s^1 \xrightarrow{b} s^2 \rvert$	.36
$\lvert s^2 \xrightarrow{a} s^1 \rvert$	.17
$\lvert s^2 \xrightarrow{b} s^1 \rvert$	.83

References

CHARNIAK, EUGENE. 1993. *Statistical Language Learning*. Cambridge: MIT Press.

HAMMOND, MICHAEL, 2003. Statistical natural language processing. U. of Arizona.

HOPCROFT, J.E., & J.D. ULLMAN. 1979. *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley.