Hidden Markov Models

464.561A Models and Technologies for Information Services

Jonghun Park

jonghun@snu.ac.kr

Dept. of Industrial Eng.

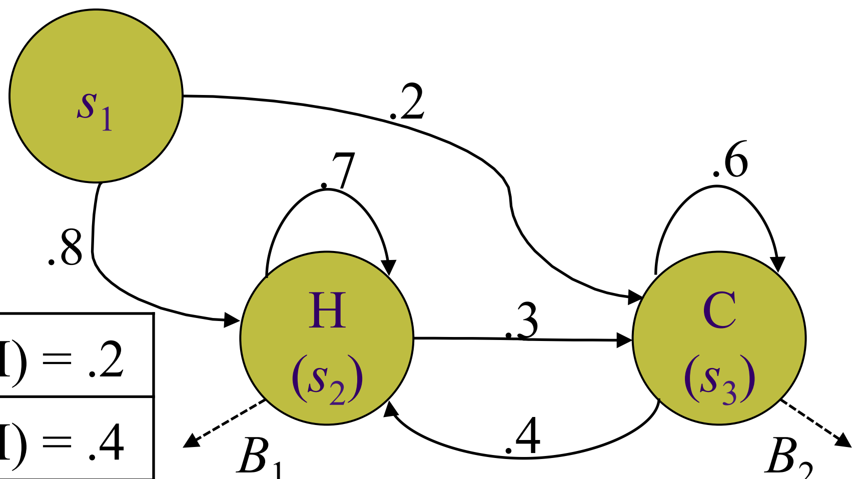
Seoul National University

3/27/11

hidden Markov model

- doubly (or bivariate) stochastic process in which an underlying stochastic process that is not observable can only be observed through another stochastic process that produces a sequence of observations
  - state process
  - observation process

a little story about HMM



$P(1\|H) = .2$
$P(2\|H) = .4$
$P(3\|H) = .4$

$P(1\|C) = .5$
$P(2\|C) = .4$
$P(3\|C) = .1$

HMM: definition

$\lambda = (S, V, A, B, \Pi)$

set of states $S = \{s_1, \ldots, s_N\}$

set of observation symbols $V = \{v_1, \ldots, v_M\}$

state transition probabilities $A = [a_{ij}]$

where $a_{ij} \triangleq P(q_{t+1} = s_j | q_t = s_i), \forall i, j = 1, \ldots, N$

$q_t, t = 1, \ldots, T$: state at time t

observation probabilities $B = [b_i(m)]$

where $b_j(m) \triangleq P(o_t = v_m | q_t = s_j), \forall i = 1, \ldots, N$

$o_t, t = 1, \ldots, T$: observation at time t

initial state probabilities $\Pi = [\pi_i]$

where $\pi_i \triangleq P(q_1 = s_i)$

stochastic constraints on HMM

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i$$

$$\sum_{j=1}^{N} \pi_i = 1, \forall i$$

assumptions

Markov assumption:

$$P(q_t|q_1 \dots q_{t-1}) = P(q_t|q_{t-1})$$

output independence:

$$P(o_t|q_1 \dots q_t \dots q_T, o_1 \dots o_t \dots o_T) = P(o_t|q_t)$$

multinomial observations:

$$P(o_t|q_t = s_j, \lambda) = \prod_{m=1}^{M} b_j(m)^{r_m^t}$$

where $r_m^t = \begin{cases} 1, & \text{if } o_t = v_m \\ 0, & \text{o.w.} \end{cases}$

3 fundamental problems

likelihood problem:
Given an HMM $\lambda = (A, B, \Pi)$ and an observation sequence O, determine the likelihood $P(O|\lambda)$

decoding problem:
Given an HMM $\lambda = (A, B, \Pi)$ and an observation sequence O, discover the best hidden state sequence Q

learning problem:
Given an an observation sequence O, the set of states S, and the set of symbols V in the HMM, learn the HMM parameters A and B

likelihood: probability evaluation

$$O \triangleq o_1 o_2 \ldots o_T$$

$$Q \triangleq q_1 q_2 \ldots q_T$$

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(o_t|q_t,\lambda)$$

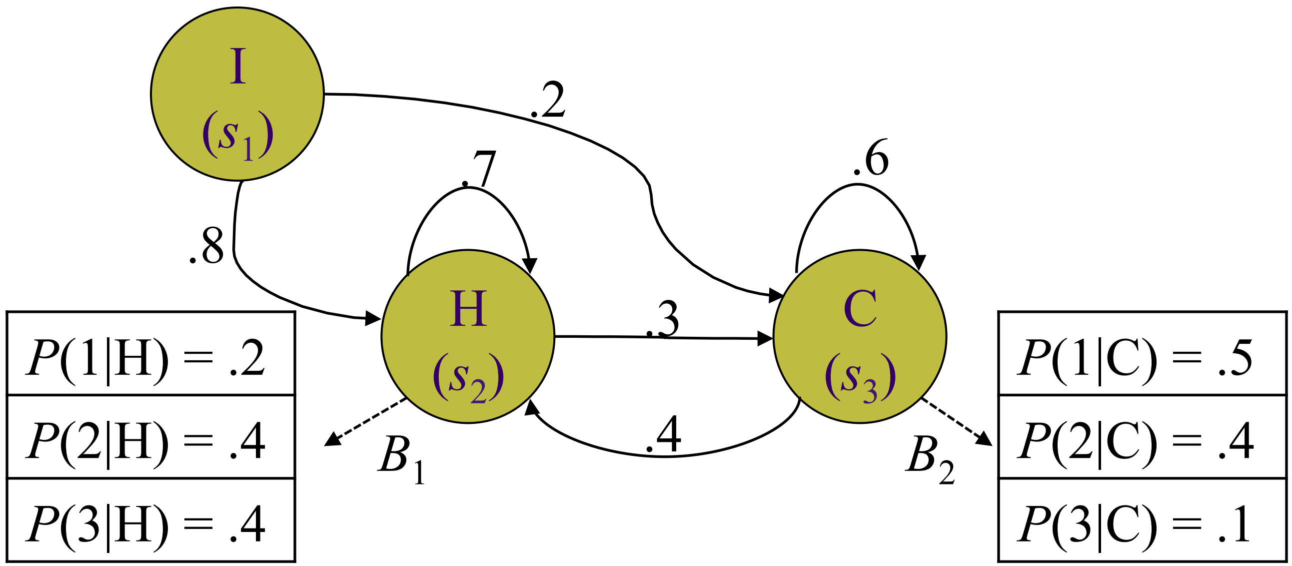
$$P(O,Q|\lambda) = P(O|Q,\lambda)P(Q|\lambda)$$

$$= \prod_{t=1}^{T} P(o_t|q_t,\lambda) \prod_{t=1}^{T} P(q_{t+1}|q_t,\lambda)$$

$$P(O|\lambda) = \sum_{Q} P(O|Q,\lambda)P(Q|\lambda)$$

$$= \sum_{q_1 q_2 \ldots q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \ldots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

likelihood computation



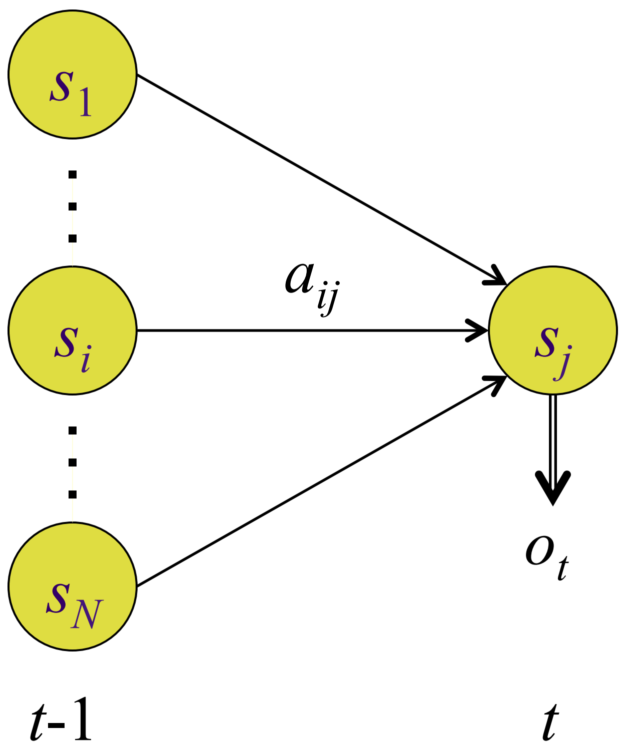
$$P(313|\lambda) = P(313, CCC|\lambda) + P(313, CCH|\lambda) + P(313, HHC|\lambda) + \ldots$$

$$P(313, HHC) = P(H|q_0) \times P(H|H) \times P(C|H) \times P(3|H) \times P(1|H) \times P(3|C)$$

likelihood: forward variable

$$\alpha_t(j) \triangleq P(o_1 o_2 \ldots o_t, q_t = j | \lambda)$$

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \right] b_j(o_t)$$



likelihood: forward algorithm

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \leq i \leq N$$

2. Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}) \qquad \begin{array}{l} 1 \leq t \leq T-1 \\ 1 \leq j \leq N \end{array}$$

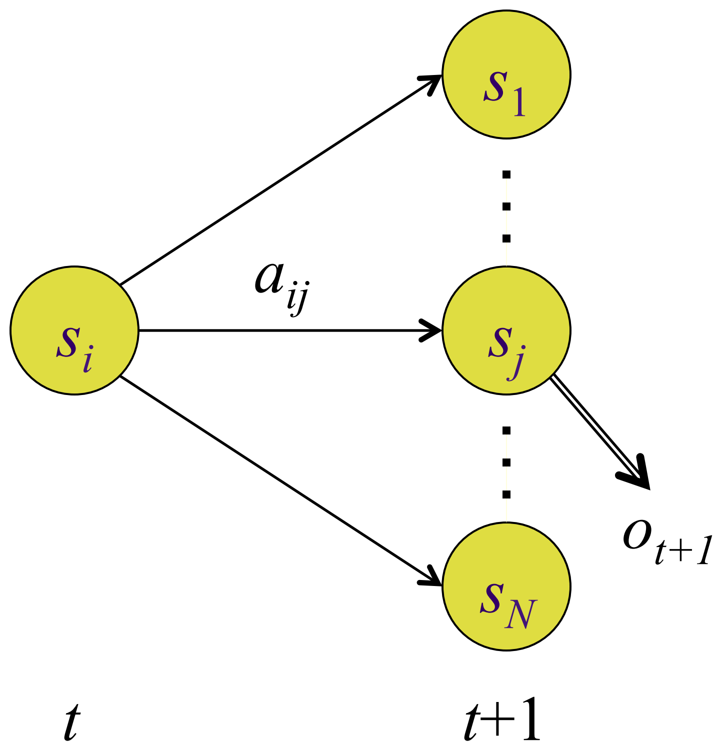
3. Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

backward variable

$$\beta_t(i) \triangleq P(o_{t+1}o_{t+2}\dots o_T | q_t = i, \lambda)$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$



backward procedure

1. initialization

$$\beta_T(i) = 1,\ 1 \leq i \leq N$$

2. induction

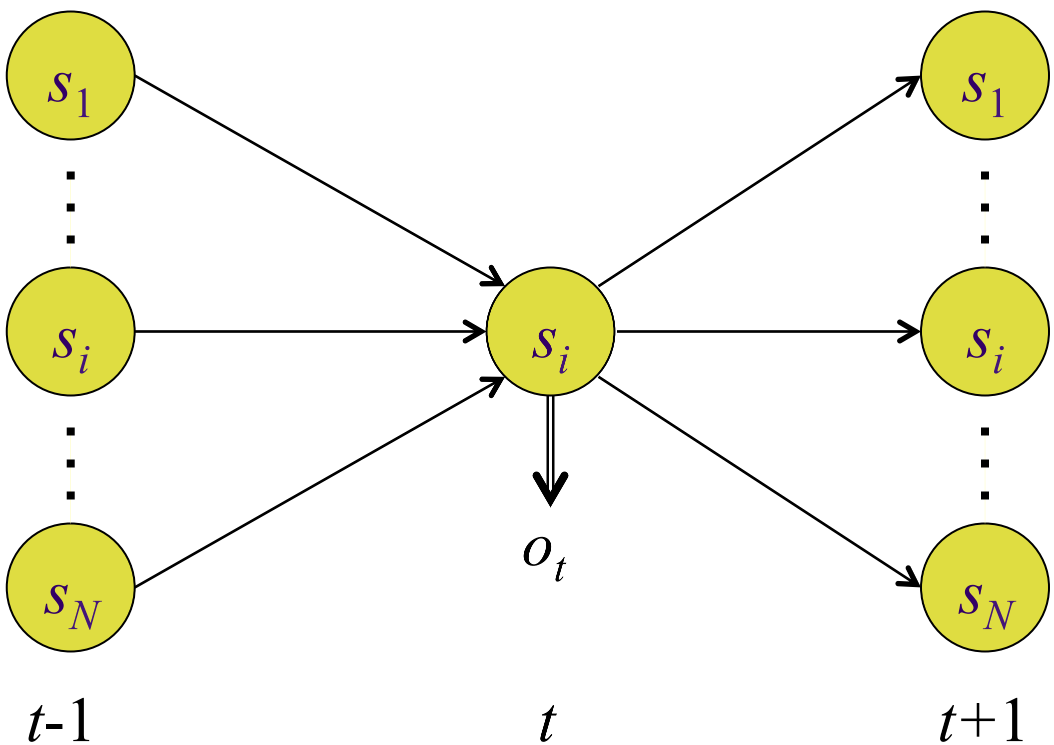
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

$$t = T-1, T-2, \ldots, 1,\ 1 \leq i \leq N$$

decoding: Viterbi algorithm

$$\gamma_t(i) \triangleq P(q_t = i | O, \lambda) = \frac{P(O, q_t = i|\lambda)}{P(O|\lambda)}$$

$$= \frac{P(O, q_t = i|\lambda)}{\sum_{i=1}^{N} P(O, q_t = i|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$



most likely state at time *t*

$$q_t^* = \arg \min_{1 \le i \le N} [\gamma_t(i)],\ 1 \le t \le T$$

Viterbi algorithm

$$\delta_t(i) \triangleq \max_{q_1 q_2 \ldots q_{t-1}} P(q_1 q_2 \ldots q_{t-1}, q_t = i, o_1 o_2 \ldots o_t | \lambda)$$

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] \cdot b_j(o_{t+1})$$

$\psi_t(j) \triangleq$ the state that maximizes $\delta_{t-1}(j)$ at time t - 1

Viterbi algorithm

1. initialization

$$\delta_1(i) = \pi_i b_i(o_1),\ \ \psi_1(i) = 0,\ \ 1 \le i \le N$$

2. recursion

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}]\, b_j(o_t)$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}],\, 2 \le t \le T,\ \ 1 \le j \le N$$

3. termination

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

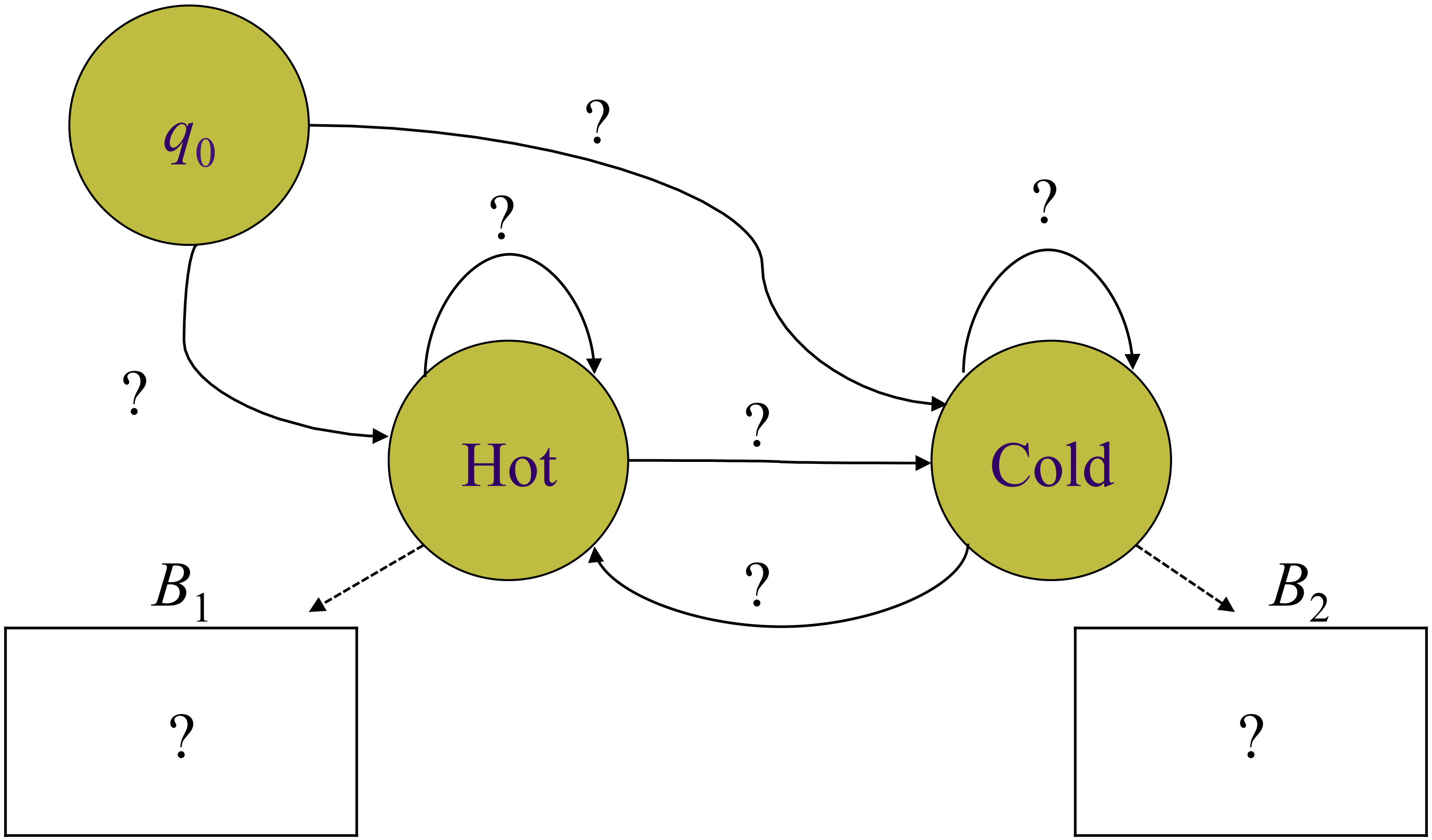
$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

4. path backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ \ t = T-1, T-2, \ldots, 1$$

learning problem

choose λ such that its likelihood $P(O|\lambda)$ is maximized



$Q = \{\text{Hot, Cold}\}$
$O = \{1, 3, 2, 2, 1, \ldots\}$

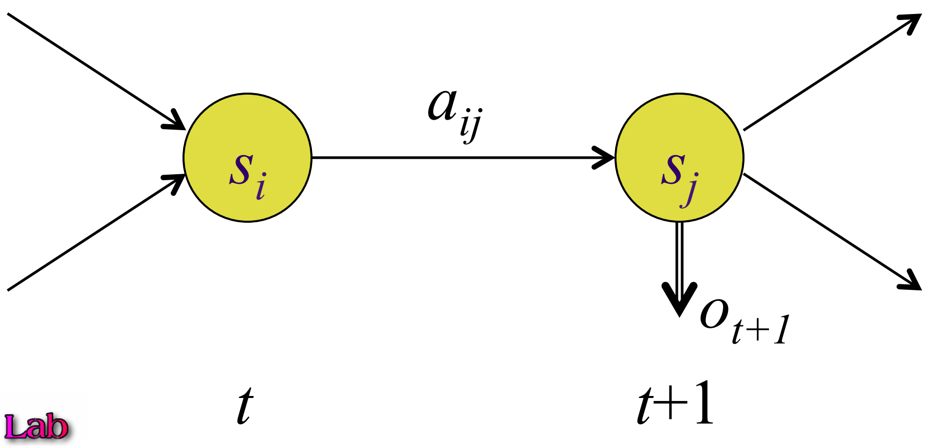
learning problem

$$\xi_t(i,j) \triangleq P(q_t = i, q_{t+1} = j | O, \lambda)$$

$$= \frac{P(q_t = i, q_{t+1} = j, O | \lambda)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}$$



learning problem

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected \# of transitions from state } i \text{ in } O$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \begin{array}{l}\text{expected \# of transitions from state } i \\ \text{to state } j \text{ in } O\end{array}$$

(re)estimation of an HMM parameters

$$\hat{\pi}_i = \text{expected frequency in state } i \text{ at } t = 1 = \gamma_1(i)$$

$$\hat{a}_{ij} = \frac{\text{expected \# of transitions from state } i \text{ to state } j}{\text{expected \# of transitions from state } i}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\hat{b}_j(m) = \frac{\text{expected \# of times in state } j \text{ and observing } v_m}{\text{expected \# of times from state } j}$$

$$= \frac{\sum_{\{t=1,\ldots,T \mid o_t = v_m\}} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

forward-backward algorithm (Baum-Welch)

initialize A, B, and Π
iterate until convergence

E-step

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}, \forall t, i$$

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}$$

$\forall t, i, j$

M-step

$$\hat{\pi}_i = \gamma_1(i)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$$

$$\hat{b}_j(m) = \frac{\sum_{t=1, \text{ s.t. } o_t = v_m}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}$$

$$\alpha_t(j) \triangleq P(o_1 o_2 \ldots o_t, q_t = j|\lambda) = \left[\sum_{i=1}^{N}\alpha_{t-1}(i)a_{ij}\right] b_j(o_t)$$

$$\beta_t(i) \triangleq P(o_{t+1}o_{t+2}\ldots o_T|q_t = i, \lambda) = \sum_{j=1}^{N} a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$

multiple training sequences

$$\mathcal{X} \triangleq \{O^k\}_{k=1}^{K}$$

$$P(\mathcal{X}|\lambda) \doteq \prod_{k=1}^{K} P(O^k|\lambda)$$

$$\hat{\pi}_i = \frac{\sum_{k=1}^{K} \gamma_1^k(i)}{K}$$

$$\hat{a}_{ij} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T^k-1} \xi_t^k(i,j)}{\sum_{k=1}^{K} \sum_{t=1}^{T^k-1} \gamma_t^k(i)}$$

$$\hat{b}_j(m) = \frac{\sum_{k=1}^{K} \sum_{\substack{t=1 \\ \text{s.t.}\, o_t^k = v_m}}^{T^k} \gamma_t^k(j)}{\sum_{k=1}^{K} \sum_{t=1}^{T^k} \gamma_t^k(j)}$$

model selection

- tuning the topology of an HMM
  - zeroing out some impossible (or unnecessary) transitions: $a_{ij} = 0$
  - moving forward only: $a_{ij} = 0$, for $j < i$
  - no big jumps: $a_{ij} = 0$, for $j > i + \tau$
- # of states
  - determined using prior information
  - can be fine-tuned by cross validation by checking the likelihood of validation sequences

finite mixture

$$P(o_t = v_m) = \sum_{j=1}^{N} P(o_t = v_m | q_t = s_j) P(q_t = s_j)$$

$$a_{ij} \doteq w_j, \ j = 1, \ldots, N, \ \forall i$$

$$\Longrightarrow P(q_t = s_j) = w_j$$

$$\Longrightarrow P(o_t = v_m) = \sum_{j=1}^{N} b_j(m) \cdot w_j$$

continuous observation densities

$$P(o_t|q_t = s_j) \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

M-step equations:

$$\hat{\mu}_j = \frac{\sum_t \gamma_t(j) o_t}{\sum_t \gamma_t(j)}$$

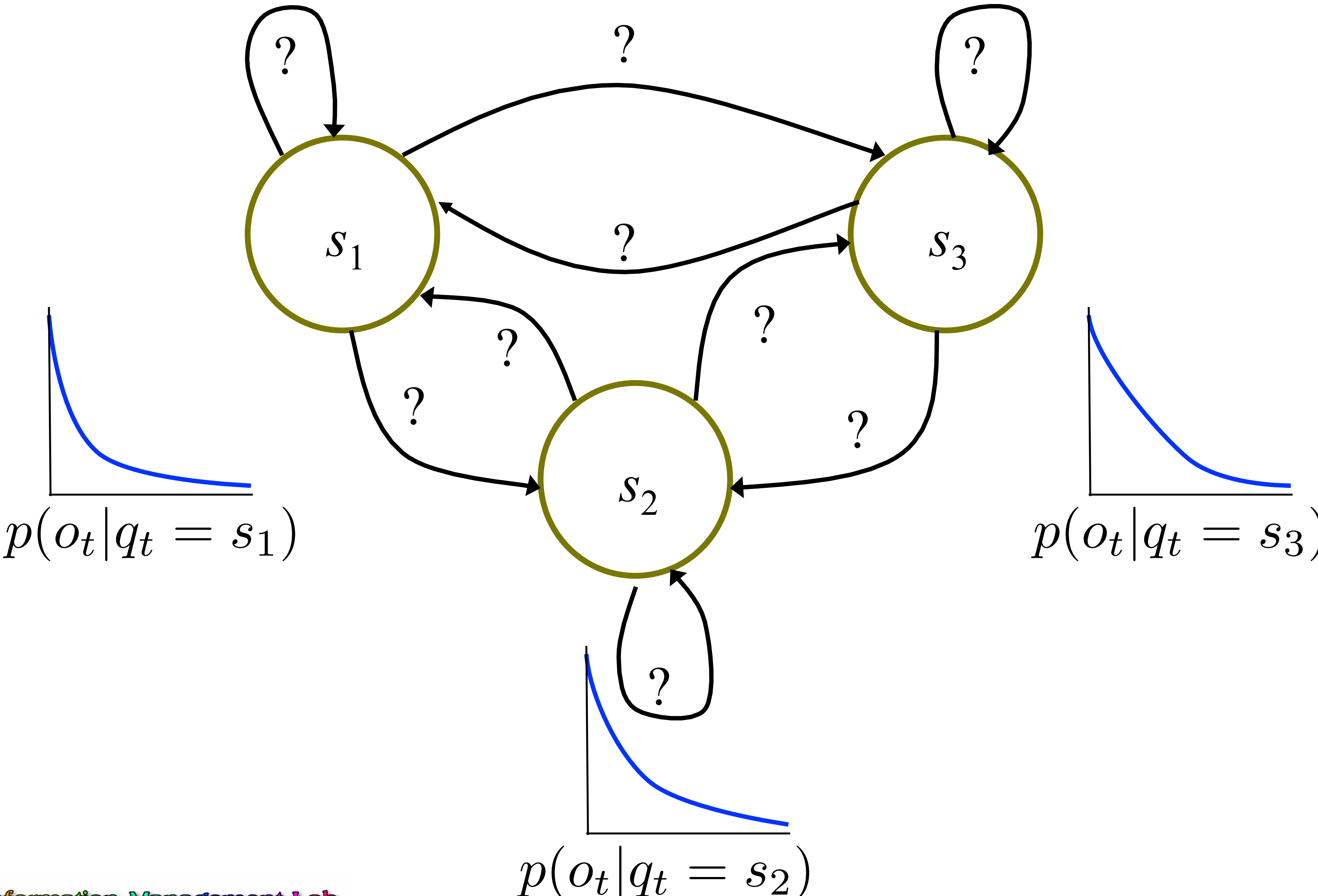
$$\hat{\sigma}_j^2 = \frac{\sum_t \gamma_t(j)(o_t - \hat{\mu}_j)^2}{\sum_t \gamma_t(j)}$$

earthquake example

(source: D. W. Chambers, et al., Hidden Markov model forecasting of earthquakes)

- $T = 1227$ earthquakes in southern California in 1932 - 2004
- (s_1, s_2, s_3) = (short, moderate, long) time between earthquakes
- $o_1, o_2, \ldots, o_T$ are the actual inter-event times
  - o_t: # days between earthquakes t-1 and t
- given $q_t = s_i$, o_t follows an exponential distribution with mean θ_i s.t. $\theta_1 < \theta_2 < \theta_3$

earthquake example



earthquake: forecasting problem

Given an HMM with parameters (A, θ_1, θ_2, θ_3, Π), and the observations up to the present, find the forecast density:

$$p(o_{t+1} = x | o_1 o_2 \ldots o_t, \lambda)$$

$$= \sum_{i=1}^{3} p(o_{t+1} = x, q_{t+1} = s_i | o_1 o_2 \ldots o_t, \lambda)$$

$$= \sum_{i=1}^{3} p(o_{t+1} = x | q_{t+1} = s_i, o_1 \ldots o_t, \lambda) \cdot p(q_{t+1} = s_i | o_1 \ldots o_t, \lambda)$$

$$= \sum_{i=1}^{3} \frac{1}{\theta_i} e^{-\frac{x}{\theta_i}} \cdot p(q_{t+1} = s_i | o_1 \ldots o_t, \lambda)$$

$$= \sum_{i=1}^{3} \frac{1}{\theta_i} e^{-\frac{x}{\theta_i}} \cdot \frac{p(q_{t+1} = s_i, o_1 \ldots o_t | \lambda)}{p(o_1 \ldots o_t | \lambda)} = \sum_{i=1}^{3} \frac{1}{\theta_i} e^{-\frac{x}{\theta_i}} \cdot \frac{\sum_{j=1}^{3} \alpha_t(j) \cdot a_{ji}}{\sum_{j=1}^{3} \alpha_t(j)}$$

earthquake: forecasting problem

$$P(o_{t+1} \leq x | o_1 o_2 \ldots o_t, \lambda)$$

$$= \sum_{i=1}^{3} (1 - e^{-\frac{x}{\theta_i}}) \cdot \frac{\sum_{j=1}^{3} \alpha_t(j) \cdot a_{ji}}{\sum_{j=1}^{3} \alpha_t(j)}$$

earthquake: the results

- $\theta_1 = 1.3$, $\theta_2 = 17.42$, $\theta_3 = 27{,}92$
- 1226 forecasts were made to find the probability of another earthquake within 7 days
- proportion of times an earthquake did actually occur within 7 days:

	[.28, .32)	[.32, .36)	[.36, .50)	[.50, 1]
proportion	.292	.300	.421	.625

classification

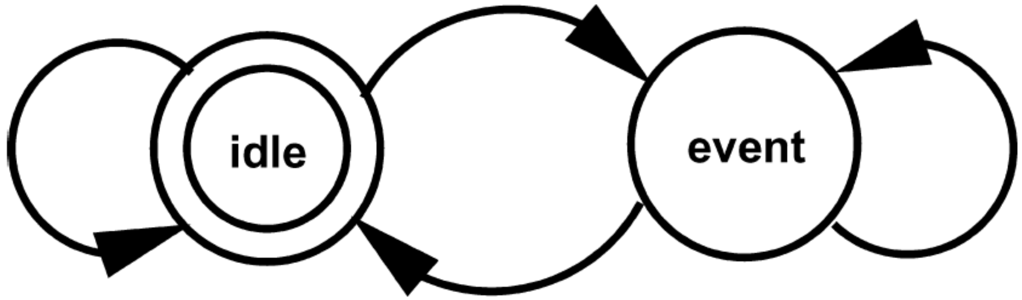
- set of HMMs, each one modeling the sequences belonging to one class

$$\arg\max_i P(\lambda_i|O) = \frac{P(O|\lambda_i)P(\lambda_i)}{\sum_j P(O|\lambda_j)P(\lambda_j)}$$

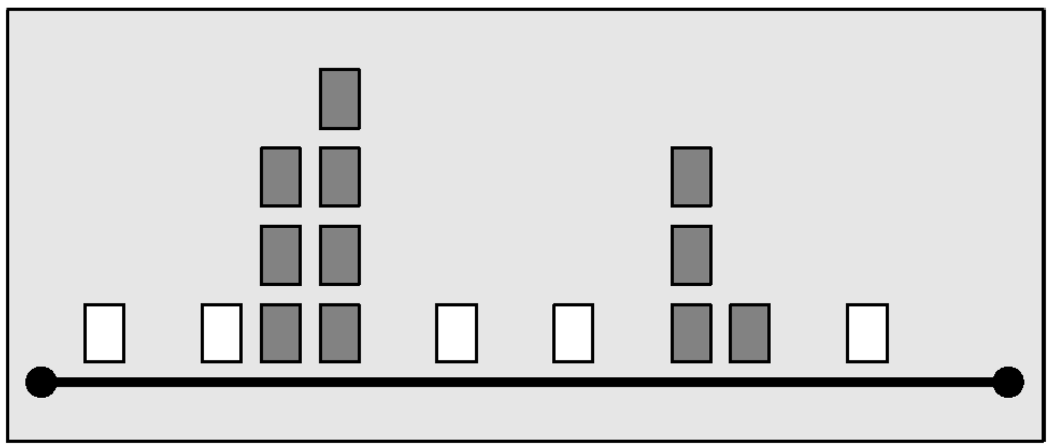
numerical issues

- scaling α and β
  - for large t, $\alpha_t(i)$ computation will exceed the precision range (even in double precision)
- smoothing
  - symbol v_m that does not appear in the training sequence will make $b_j(m) = 0$
- imbalance between emission and transition probabilities
  - $b_j(m) << a_{ij}$
  - $P(O|\lambda)$ becomes mostly influenced by $b_j(m)$

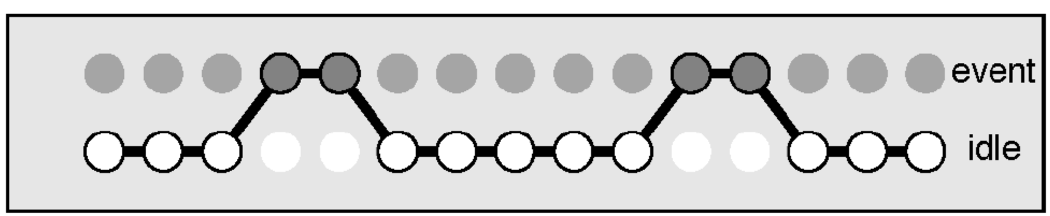
HMM for Hot Topic Detection



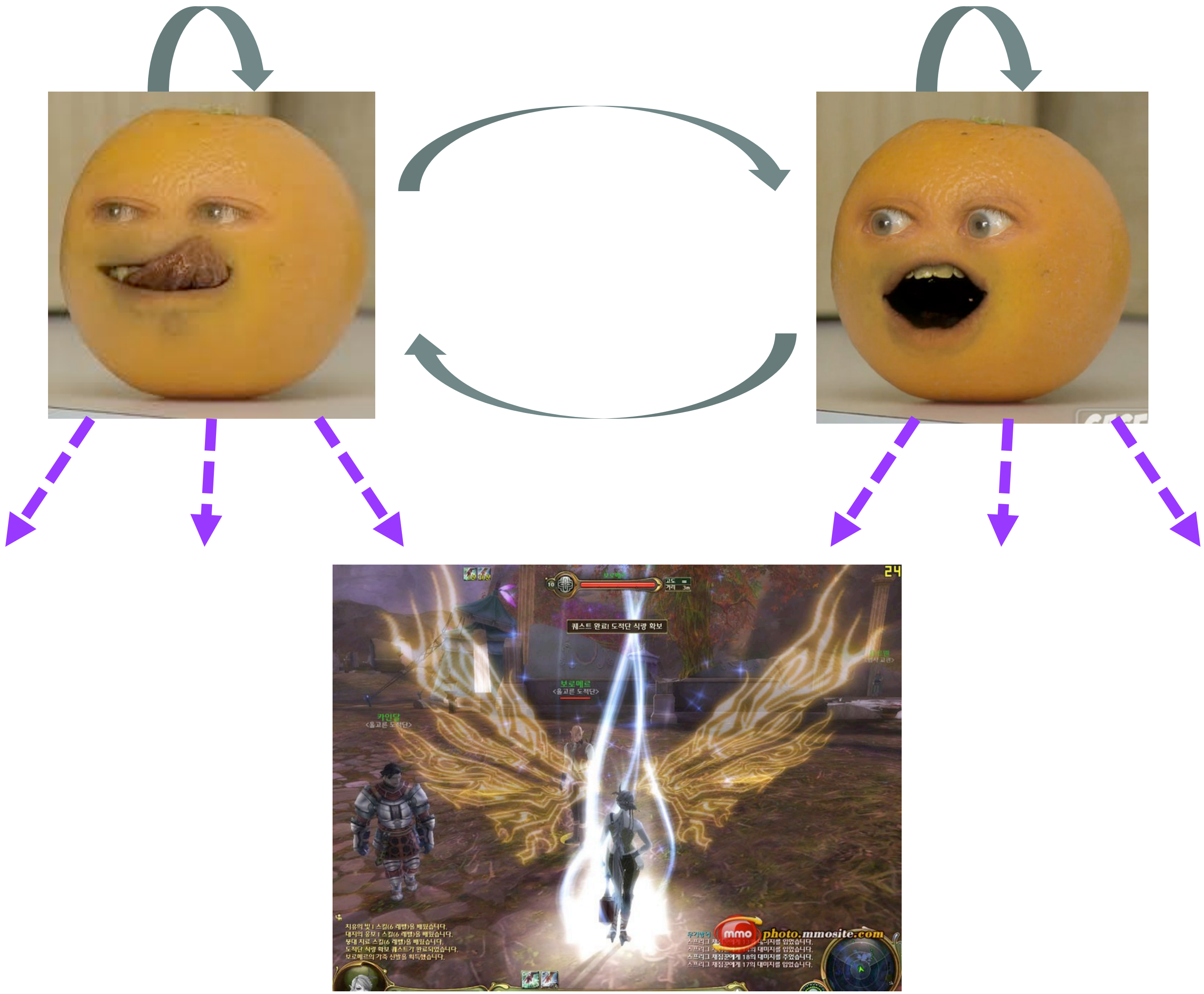
df_{i}



state



user state mining for games



automatic composition of Mozart style music

readings

- K. Seymore, A. McCallum, and R. Rosenfeld, "Learning Hidden Markov Model Structure for Information Extraction," *Proc. AAAI Workshop on Machine Learning for Information Extraction*, 1999.
- A. Srivastava, et al., "Credit Card Fraud Detection Using Hidden Markov Model," *IEEE Trans. Dependable and Secure Computing*, vol. 5, no. 1, pp. 37-48, 2008.
- S. M. Siddiqi and A. W. Moore, "Fast Inference and Learning in Large-State-Space HMMs," *Proc. 22nd Int'l Conf. Machine Learning*, 2005.
- O. Netzer, J. M. Lattin, and V. Srinivasan, "A Hidden Markov Model of Customer Relationship Dynamics," *Marketing Science*, vol. 27, no. 2, pp. 185-204, 2008.
- M. J. Zaki, C. D. Carothers, and B. K. Szymanski, "VOGUE: A Variable Order Hidden Markov Model with Duration Based on Frequent Sequence Mining," *ACM Trans. Knowledge Discovery from Data*, vol. 4, no. 1, 2010.
- J. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Proc. 8th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2002.

references

- E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, 2004.
- O. C. Ibe, *Markov Processes for Stochastic Modeling*, Academic Press, 2009.
- F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, 1997.
- D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd Ed., Prentice Hall, 2009
- T. Koski, *Hidden Markov Models for Bioinformatics*, Kluwer Academic Publishers, 2001.
- L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- W. Zucchini and I. L. MacDonald, *Hidden Makov Models for Time Series*: *An Introduction Using R*, CRC Press, 2009.