E9 261

07-03-2016

Recap …

- Gaussian Mixture Models
  - Static modeling
- Dynamic Time Warping
  - Non-statistical Modeling
- Hidden Markov Modeling
  - Statistical and sequence model

Three basic problems for HMMs

Evaluation Given the observation sequence $O = O_1 O_2 \ldots O_T$ and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, i.e., the probability of the observation sequence given the model

Recognition Given the observation sequence $O = O_1 O_2 \ldots O_T$ and a model $\lambda = (A, B, \pi)$, how do we choose a corresponding state sequence $Q = q_1 q_2 \ldots q_T$ which is optimal in some sense, i.e., best explains the observations

Training Given the observation sequence $O = O_1 O_2 \ldots O_T$, how do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$

Forward procedure

- We define a forward variable $\alpha_j(t)$ as the probability of the partial observation seq. until time t, with state S_j at time t

$$\alpha_j(t) = P(O_1 O_2 ... O_t, q_t = S_j | \lambda)$$

- This can be computed inductively

$$\alpha_j(1) = \pi_j b_{jO_1} \qquad 1 \leq j \leq N$$

$$\alpha_j(t+1) = \Big(\sum_{i=1}^{N} \alpha_i(t) a_{ij}\Big) b_{jO_{t+1}} \qquad 1 \leq t \leq T-1$$

- Then with $N^2 T$ operations:

$$P(O|\lambda) = \sum_{i=1}^{N} P(O, q_T = S_i | \lambda) = \sum_{i=1}^{N} \alpha_i(T)$$

Viterbi algorithm

- Finding the best single sequence means computing $\text{argmax}_Q P(Q|O,\lambda)$, equivalent to $\text{argmax}_Q P(Q,O|\lambda)$
- The Viterbi algorithm (dynamic programming) defines $\delta_j(t)$, i.e., the highest probability of a single path of length t which accounts for the observations and ends in state S_j

$$\delta_j(t) = \max_{q_1,q_2,\ldots,q_{t-1}} P(q_1 q_2 \ldots q_t = j, O_1 O_2 \ldots O_t|\lambda)$$

- By induction

$$\begin{aligned} \delta_j(1) &= \pi_j b_{jO_1} & 1 \le j \le N \\ \delta_j(t+1) &= \left(\max_i \delta_i(t) a_{ij}\right) b_{jO_{t+1}} & 1 \le t \le T-1 \end{aligned}$$

- With backtracking (keeping the maximizing argument for each t and j) we find the optimal solution

Baum-Welch Reestimation

- Reestimation formulas

$$\bar{\pi}_i = \gamma_i(1) \qquad \bar{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \qquad \bar{b_{jk}} = \frac{\sum_{O_t = v_k} \gamma_j(t)}{\sum_{t=1}^{T} \gamma_j(t)}$$

- Baum et al. proved that if current model is $\lambda = (A, B, \pi)$ and we use the above to compute $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ then either
  - $\bar{\lambda} = \lambda$ – we are in a critical point of the likelihood function
  - $P(O|\bar{\lambda}) > P(O|\lambda)$ – model $\bar{\lambda}$ is more likely
- If we iteratively reestimate the parameters we obtain a maximum likelihood estimate of the HMM
- Unfortunately this finds a local maximum and the surface can be very complex

For Gaussian mixtures, we define the probability that the ℓ^{th} component of the i^{th} mixture generated observation o_t as

$$\gamma_{i\ell}(t) = \gamma_i(t)\frac{c_{i\ell}b_{i\ell}(o_t)}{b_i(o_t)} = p(Q_t = i, X_{it} = \ell | O, \lambda)$$

where X_{it} is a random variable indicating the mixture component at time t for state i.

From the previous section on Gaussian Mixtures, we might guess that the update equations for this case are:

$$c_{i\ell} = \frac{\sum_{t=1}^{T} \gamma_{i\ell}(t)}{\sum_{t=1}^{T} \gamma_i(t)}$$

$$\mu_{i\ell} = \frac{\sum_{t=1}^{T} \gamma_{i\ell}(t) o_t}{\sum_{t=1}^{T} \gamma_{i\ell}(t)}$$

$$\Sigma_{i\ell} = \frac{\sum_{t=1}^{T} \gamma_{i\ell}(t)(o_t - \mu_{i\ell})(o_t - \mu_{i\ell})^T}{\sum_{t=1}^{T} \gamma_{i\ell}(t)}$$

For Multiple Observation Sequences

$$\pi_i = \frac{\sum_{e=1}^{E} \gamma_i^e(1)}{E}$$

$$c_{i\ell} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t)}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_i^e(t)}$$

$$\mu_{i\ell} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t) o_t^e}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t)}$$

Non-ergodic HMMs

- Until now we have only considered ergodic (fully connected) HMMs
  - every state can be reached from any state in a finite number of steps

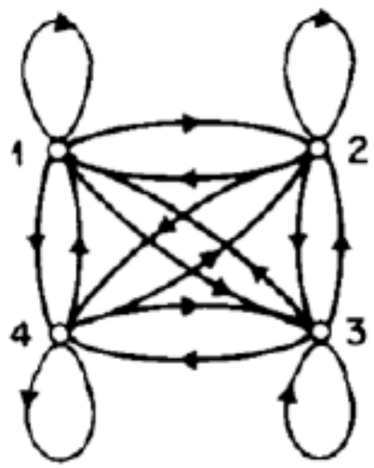

Figure: Ergodic HMM

- Left-right (Bakis) model good for speech recognition
  - as time increases the state index increases or stays the same
  - can be extended to parallel left-right models

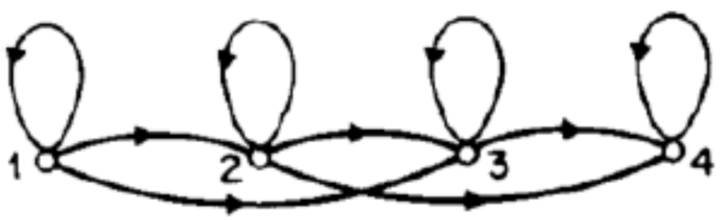

Figure: Left-right HMM

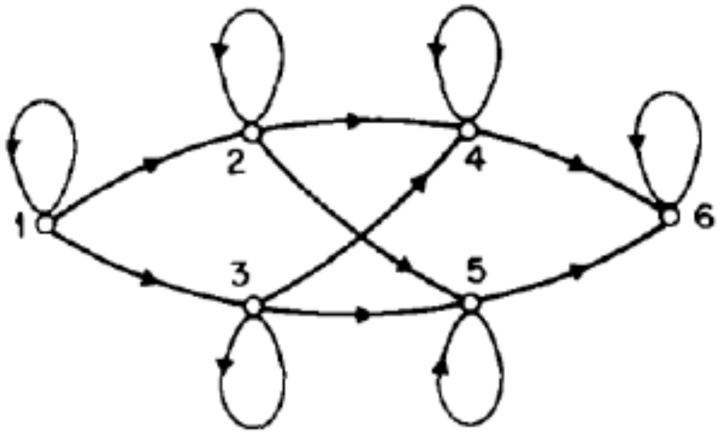

Figure: Parallel HMM

Other Considerations

- Implementation issues
  - Scaling
- Initialization
- Amount of Training Data
- Model complexity
- Whole word based Recognition System