AUTOMATIC SPEECH RECOGNITION

Supplementary No.1 for Lecture 9

Three Basic Problems for HMMs

Topics

- The three basic problems for HMMs
- The forward-backward Procedure
- The Viterbi Algorithm
- The Baum-Welch Re-estimation Procedure
- Speech recognizer using HMMs

Continuous-density HMMs

- At this point we will consider observations which are discrete symbols only. (Easier to understand)
- Feature vectors with continuous values can be converted to symbols using “Vector Quantization (VQ)”.
- VQ will be discussed later in this course.

Three Basic Problems for HMMs

- Problem1: Evaluation Problem

  Given the observation sequence $\boldsymbol{O} = \{O_1, O_2, O_3, ..., O_n\}$ and the model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$, how can the observation sequence probability $P(O|\lambda)$ be computed?
- Problem2: Hidden State Sequence Uncovering

  Given the observation sequence $\boldsymbol{O} = \{O_1, O_2, O_3, ..., O_n\}$ and the model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$, how can a state sequence $Q = \{q1, q2, ..., qT\}$, which is optimal in some sense, be chosen?
- Problem3: Training Problem

  Given example observations $\boldsymbol{O}$, how can the model parameters $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$ be adjusted to maximize the observation sequence probability $P(O|\lambda)$?

Three Basic Problems for HMMs

- Problem1: Evaluation Problem
- Given the observation sequence $\boldsymbol{O} = \{O_1, O_2, O_3, \ldots, O_n\}$ and the model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$, how can the observation sequence probability $P(O|\lambda)$ be computed?
- Solution is used for the scoring of each word model based on the given O for recognizing an unknown word.
- → "The Forward-Backward Algorithm"

Direct Probability Evaluation

Given the observation sequence $\boldsymbol{O} = \{O_1, O_2, O_3, \ldots, O_n\}$ and the model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$, how can the observation sequence probability $P(O|\lambda)$ be computed?

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

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

consider the fixed state sequence: $Q = q_1 q_2 \ldots q_T$

$$P(O|Q,\lambda) = b_{q1}(o_1) b_{q2}(o_2) \ldots b_{qT}(o_T)$$

$$P(Q|\lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \ldots a_{qT-1qT}$$

Therefore:

$$P(O|\lambda) = \sum_{q1,q2,\ldots,qT} \pi_{q1} b_{q1}(o_1) a_{q1q2} b_{q2}(o_2) \ldots a_{qT-1qT} b_{qT}(o_T)$$

Calculation required: ≈ 2T x N^T

MIT OCW

The Forward Procedure

- Define: the forward variable, $\alpha_t(i)$, as the probability of the partial observation sequence up to time t and being at state s_i at time t, given the model λ

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = s_i | \lambda)$$

- It can be shown that

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N \quad \text{and} \quad P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

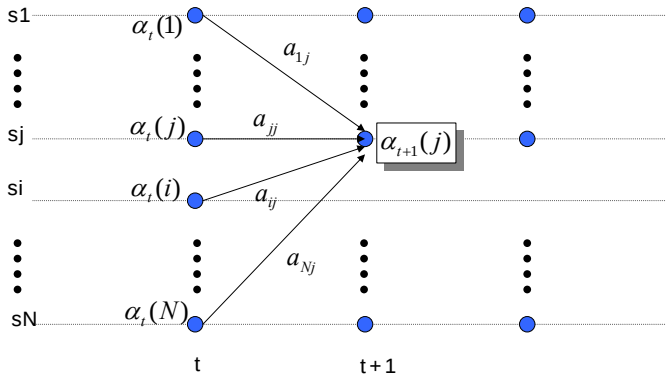
- By induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1}), \quad 1 \le t \le T-1, 1 \le j \le N$$

Calculation required: ≈ N^2 T

The Forward Procedure

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1}), \quad 1 \le t \le T-1, 1 \le j \le N$$



The Backward Procedure

- Define: the backward variable, $\beta_t(i)$, as the probability of the partial observation sequence from time $t+1$ to the end, given state s_i at time t and the model λ

$$\beta_t(i) = P(o_{t+1}o_{t+2}...o_T, q_t = s_i \mid \lambda)$$

- It can be shown that

$$\beta_T(i) = 1, \quad 1 \le i \le N \quad \text{and} \quad P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i)$$

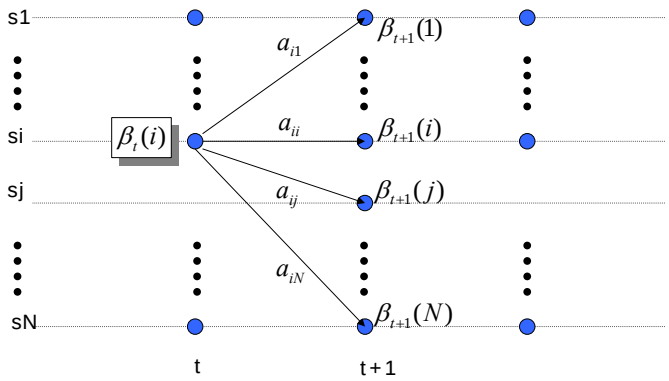
- By induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1})\beta_{t+1}(j), \qquad t = T-1, T-2, ..., 1, 1 \le i \le N$$

Calculation required: $\approx N^2 T$

The Backward Procedure

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1})\beta_{t+1}(j), \qquad t = T-1, T-2, ..., 1, 1 \le i \le N$$



Three Basic Problems for HMMs

- Problem2: Hidden State Sequence Uncovering
- Given the observation sequence $\boldsymbol{O} = \{O_1, O_2, O_3, ..., O_n\}$ and the model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$, how can a state sequence $Q = \{q1, q2, ..., qT\}$, which is optimal in some sense, be chosen?
- Solution is used to develop an understanding of the physical meaning of the model states.
- → "The Viterbi Algorithm"

Hidden State Uncovering

- Given the observation sequence $\boldsymbol{O} = \{O_1, O_2, O_3, ..., O_n\}$ and the model $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$, how can a state sequence $Q = \{q1, q2, ..., qT\}$, which is optimal in some sense, be chosen?

Possible optimal criteria (1 or 2)

1. Choose states, q_t, which are individually most likely. This maximizes the expected number of correct individual states. (Optimal state sequence may not obey state transition constraints.)
2. Choose the single best state sequence which maximizes $P(O,Q \mid \lambda)$.

Hidden State Uncovering

Choose states, q_t, which are individually most likely

- Define $\gamma_t(i)$, as the probability of being at state s_i at time t, given the observation sequence O and the model λ

$$\gamma_t(i) = P(q_t = s_i \mid O, \lambda)$$

- It can be shown that

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)}$$

- where the denominator ensures that $\sum_{i=1}^{N} \gamma_t(i) = 1, \forall t$
- Then the individually most likely state, q_t, at time t is:

$$q_t = \arg\max_{1 \le i \le N} [\gamma_t(i)], \qquad 1 \le t \le T$$

Hidden State Uncovering

Choose the single best state sequence

- This can be done by using the Viterbi Algorithm.
- Define $\delta_t(i)$, as the highest probability along a single path, at time t, which accounts for the first t observations, i.e.

$$\delta_t(i) = \max_{q1,q2,\ldots,qt-1} P(q_1 q_2 \ldots q_{t-1}, q_t = s_i, o_1 o_2, \ldots o_t \mid \lambda)$$

- By induction

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

- To retrieve the state sequence, we must keep track of the state sequence which gave the best path, at time t, to state s_i.

The Viterbi Algorithm

1. Initialization:

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

2. Recursion:

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

3. Termination:

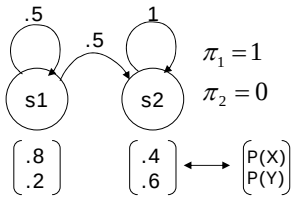
$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$
$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

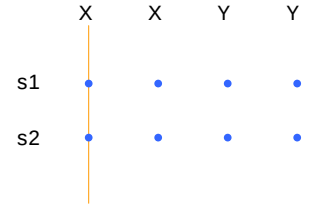
4. Path (state sequence) backtracking:

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

MIT OCW

Example





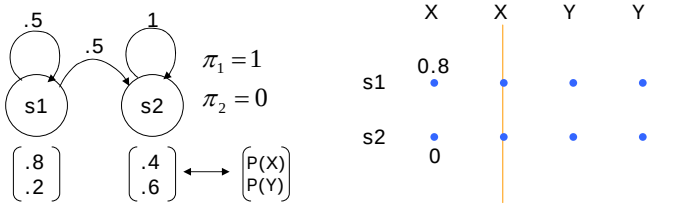
O = {X,X,Y,Y}

t=1

$$\delta_1(1) = \pi_1 b_1(X) = 1 \times .8 = .8$$
$$\delta_1(2) = \pi_2 b_2(X) = 0 \times .4 = 0$$
$$\psi_1(1) = \psi_1(2) = 0$$

initialization

Example



O = {X,X,Y,Y}

t=2

$$\delta_2(1) = \max_{1\le i\le N}[\delta_1(i)a_{i1}]b_1(X) = \max(\{.8\times.5\times.8,0\times0\times.8\}) = 0.32$$

$$\delta_2(2) = \max_{1\le i\le N}[\delta_1(i)a_{i2}]b_2(X) = \max(\{.8\times.5\times.4,0\times.4\times.4\}) = 0.16$$

$$\psi_2(1) = \arg\max_{1\le i\le N}[\delta_1(i)a_{i1}] = 1$$

$$\psi_2(2) = \arg\max_{1\le i\le N}[\delta_1(i)a_{i2}] = 1$$

Example

O = {X,X,Y,Y}

t=3

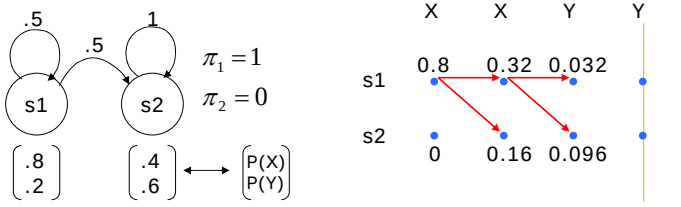
$$\delta_3(1) = \max_{1\le i\le N}[\delta_2(i)a_{i1}]b_1(Y) = \max(\{.32\times.5\times.2,.16\times0\times.2\}) = 0.032$$

$$\delta_3(2) = \max_{1\le i\le N}[\delta_2(i)a_{i2}]b_2(Y) = \max(\{.32\times.5\times.6,.16\times.4\times.6\}) = 0.096$$

$$\psi_3(1) = \arg\max_{1\le i\le N}[\delta_2(i)a_{i1}] = 1$$

$$\psi_3(2) = \arg\max_{1\le i\le N}[\delta_2(i)a_{i2}] = 1$$

Example



O = {X,X,Y,Y}

t=4

$$\delta_4(1) = \max_{1\le i\le N}[\delta_3(i)a_{i1}]b_1(Y) = \max(\{.032\times.5\times.2,.096\times0\times.2\}) = 0.0032$$

$$\delta_4(2) = \max_{1\le i\le N}[\delta_3(i)a_{i2}]b_2(Y) = \max(\{.032\times.5\times.6,.096\times1\times.6\}) = 0.0576$$

$$\psi_4(1) = \arg\max_{1\le i\le N}[\delta_3(i)a_{i1}] = 1$$

$$\psi_4(2) = \arg\max_{1\le i\le N}[\delta_3(i)a_{i2}] = 2$$

Example

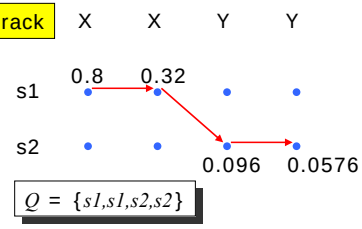
O = {X,X,Y,Y}

terminate

$$P^* = \max_{1\le i\le N}[\delta_4(i)] = 0.0576$$

$$q_T^* = \arg\max_{1\le i\le N}[\delta_4(i)] = 2$$

Backtrack



$Q = \{s1,s1,s2,s2\}$

Three Basic Problems for HMMs

- Problem3: Training Problem
- Given example observations $\boldsymbol{O}$, how can the model parameters $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$ be adjusted to maximize the observation sequence probability $P(O|\lambda)$?
- Solution is used to optimally obtain model parameters for each word model using training utterances.
- → "The Buam-Welch Re-estimation Procedures"

Baum-Welch Re-estimating Procedures

- Define $\xi_t(i,j)$, as the probability of being in state s_i at time t and state s_j at time $t+1$,given the model λ and the observation sequence

$$\zeta_t(i,j) = P(q_t = s_i, q_t = s_j \mid O, \lambda)$$

- then

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

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

Baum-Welch Re-estimating Procedures

- Summing $\gamma_t(i,j)$, and $\xi_t(i,j)$, we get:

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from } s_i$$

$$\sum_{t=1}^{T-1} \zeta_t(i,j) = \text{expected number of transitions from } s_i \text{ to } s_j$$

Baum-Welch Re-estimating Procedures

$$\bar{\pi} = \text{expected number of times in state } s_i \text{ at } t = 1$$

$$\bar{\pi} = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } s_i \text{ to } s_j}{\text{expected number of transitions from state } s_i}$$

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

Baum-Welch Re-estimating Procedures

$$\bar{b}_j(k) = \frac{\text{expected number of times in } s_j \text{ with symbol } v_k}{\text{expected number of times in } s_j}$$

$$\bar{b}_j(k) = \frac{\sum_{t=1, ot=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

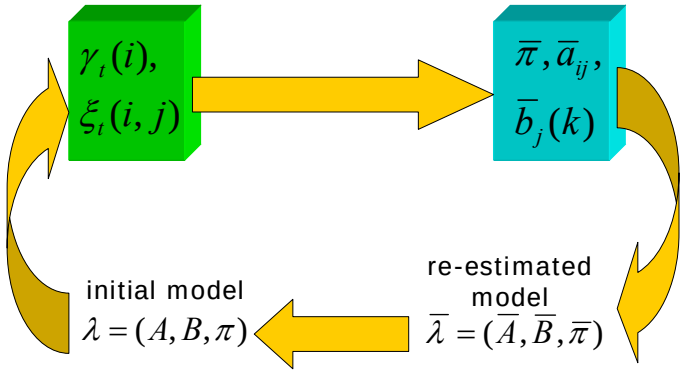
Baum-Welch Re-estimating Procedures

- if $\lambda = (A, B, \pi)$ is the initial model, and $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ is the re-estimated model, then we can be proven that either:

1. The initial model, λ, defines a critical point of the likelihood function, in which case $\bar{\lambda} = \lambda$, or
2. Model $\bar{\lambda}$ is more likely than λ, in the sense that $P(O|\bar{\lambda}) > P(O|\lambda)$

So we can improve the probability of O being observed from the model if we iteratively use $\bar{\lambda}$ in place of λ and repeat the re-estimation until some limiting point. The resulting model is called “the maximum likelihood HMM”

Baum-Welch Re-estimating Procedures



Speech Recognizer Based on HMMs