CS578- Speech Signal Processing
Lecture 11: Hidden Markov Models, HMM

Yannis Stylianou

University of Crete, Computer Science Dept., Multimedia Informatics Lab
yannis@csd.uoc.gr

Univ. of Crete

Outline

Formulas and definitions

- a *Markov chain* or *process* is a sequence of events, usually called *states*, $Q = \{q_1, \cdots q_K\}$), the probability of each of which is dependent only on the event immediately preceding it.
- a *Hidden Markov Model* (HMM) represents stochastic sequences as Markov chains where the states are not directly observed, but are associated with a probability density function (pdf)

Formulas and definitions

- a *Markov chain* or *process* is a sequence of events, usually called *states*, $Q = \{q_1, \cdots q_K\}$), the probability of each of which is dependent only on the event immediately preceding it.
- a *Hidden Markov Model* (HMM) represents stochastic sequences as Markov chains where the states are not directly observed, but are associated with a probability density function (pdf)

- The generation of a random sequence in HMM is the result of a random walk in the chain (i.e. the browsing of a random sequence of states $Q = \{q_1, \cdots q_K\}$) and of a draw (called an *emission*) at each visit of a state.
- In pattern recognition (and speech recognition) with HMMs, we are interested to associate a sequence of states $Q = \{q_1, \cdots q_K\}$ to a sequence of observations $X = \{x_1, \cdots x_K\}$).
- The true sequence of states is therefore *hidden* by a first layer of stochastic processes.

- The generation of a random sequence in HMM is the result of a random walk in the chain (i.e. the browsing of a random sequence of states $Q = \{q_1, \cdots q_K\}$) and of a draw (called an *emission*) at each visit of a state.
- In pattern recognition (and speech recognition) with HMMs, we are interested to associate a sequence of states $Q = \{q_1, \cdots q_K\}$ to a sequence of observations $X = \{x_1, \cdots x_K\}$).
- The true sequence of states is therefore *hidden* by a first layer of stochastic processes.

- The generation of a random sequence in HMM is the result of a random walk in the chain (i.e. the browsing of a random sequence of states $Q = \{q_1, \cdots q_K\}$) and of a draw (called an *emission*) at each visit of a state.
- In pattern recognition (and speech recognition) with HMMs, we are interested to associate a sequence of states $Q = \{q_1, \cdots q_K\}$ to a sequence of observations $X = \{x_1, \cdots x_K\}$).
- The true sequence of states is therefore *hidden* by a first layer of stochastic processes.

HMM terminology

- *Emission probabilities*: are the pdfs (usually Gaussians) that characterize each state q_i, i.e. $p(x|q_i)$. To simplify the notations, they will be denoted $b_i(x)$.
- *Transition probabilities*: are the probability to go from a state i to a state j, i.e. $P(q_j|q_i)$. They are stored in matrices where each term a_{ij} denotes a probability $P(q_j|q_i)$.
- *Non-emitting initial and final states*: For a finite length random sequence, two additional states are used in order to model the "start" or "end" events. These states are not associated with some emission probabilities.
- *Initial state distribution* $P(I|q_j)$: Transitions starting from the initial state.
- *Final-absorbent state*: The final state usually has only one non-null transition that loops onto itself with a probability of 1

HMM TERMINOLOGY

- *Emission probabilities*: are the pdfs (usually Gaussians) that characterize each state q_i, i.e. $p(x|q_i)$. To simplify the notations, they will be denoted $b_i(x)$.
- *Transition probabilities*: are the probability to go from a state i to a state j, i.e. $P(q_j|q_i)$. They are stored in matrices where each term a_{ij} denotes a probability $P(q_j|q_i)$.
- *Non-emitting initial and final states*: For a finite length random sequence, two additional states are used in order to model the "start" or "end" events. These states are not associated with some emission probabilities.
- *Initial state distribution $P(I|q_j)$*: Transitions starting from the initial state.
- *Final-absorbent state*: The final state usually has only one non-null transition that loops onto itself with a probability of 1

HMM terminology

- *Emission probabilities*: are the pdfs (usually Gaussians) that characterize each state q_i, i.e. $p(x|q_i)$. To simplify the notations, they will be denoted $b_i(x)$.
- *Transition probabilities*: are the probability to go from a state i to a state j, i.e. $P(q_j|q_i)$. They are stored in matrices where each term a_{ij} denotes a probability $P(q_j|q_i)$.
- *Non-emitting initial and final states*: For a finite length random sequence, two additional states are used in order to model the "start" or "end" events. These states are not associated with some emission probabilities.
- *Initial state distribution* $P(I|q_j)$: Transitions starting from the initial state.
- *Final-absorbent state*: The final state usually has only one non-null transition that loops onto itself with a probability of 1

HMM TERMINOLOGY

- *Emission probabilities*: are the pdfs (usually Gaussians) that characterize each state q_i, i.e. $p(x|q_i)$. To simplify the notations, they will be denoted $b_i(x)$.
- *Transition probabilities*: are the probability to go from a state i to a state j, i.e. $P(q_j|q_i)$. They are stored in matrices where each term a_{ij} denotes a probability $P(q_j|q_i)$.
- *Non-emitting initial and final states*: For a finite length random sequence, two additional states are used in order to model the "start" or "end" events. These states are not associated with some emission probabilities.
- *Initial state distribution $P(I|q_j)$*: Transitions starting from the initial state.
- *Final-absorbent state*: The final state usually has only one non-null transition that loops onto itself with a probability of 1

HMM TERMINOLOGY

- *Emission probabilities*: are the pdfs (usually Gaussians) that characterize each state q_i, i.e. $p(x|q_i)$. To simplify the notations, they will be denoted $b_i(x)$.
- *Transition probabilities*: are the probability to go from a state i to a state j, i.e. $P(q_j|q_i)$. They are stored in matrices where each term a_{ij} denotes a probability $P(q_j|q_i)$.
- *Non-emitting initial and final states*: For a finite length random sequence, two additional states are used in order to model the “start” or “end” events. These states are not associated with some emission probabilities.
- *Initial state distribution $P(I|q_j)$*: Transitions starting from the initial state.
- *Final-absorbent state*: The final state usually has only one non-null transition that loops onto itself with a probability of 1

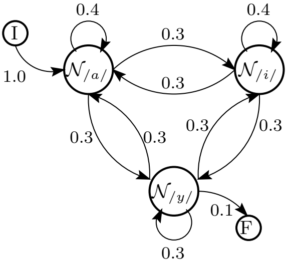
HMM terminology

- *Ergodic HMM*: an HMM allowing for transitions from any emitting state to any other emitting state
- *Left-right HMM*: an HMM where the transitions only go from one state to itself or to a unique follower.

HMM terminology

- *Ergodic HMM*: an HMM allowing for transitions from any emitting state to any other emitting state
- *Left-right HMM*: an HMM where the transitions only go from one state to itself or to a unique follower.

HMM1:

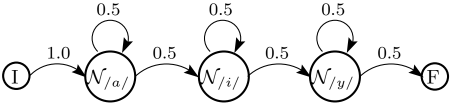


Transition matrix

$$\begin{bmatrix} 0.0 & \mathbf{1.0} & 0.0 & 0.0 & 0.0 \\ 0.0 & \mathbf{0.4} & \mathbf{0.3} & \mathbf{0.3} & 0.0 \\ 0.0 & \mathbf{0.3} & \mathbf{0.4} & \mathbf{0.3} & 0.0 \\ 0.0 & \mathbf{0.3} & \mathbf{0.3} & \mathbf{0.3} & \mathbf{0.1} \\ 0.0 & 0.0 & 0.0 & 0.0 & \mathbf{1.0} \end{bmatrix}$$

HMM EXAMPLES

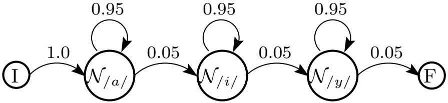
HMM2:



Transition matrix

$$\begin{bmatrix} 0.0 & \mathbf{1.0} & 0.0 & 0.0 & 0.0 \\ 0.0 & \mathbf{0.5} & \mathbf{0.5} & 0.0 & 0.0 \\ 0.0 & 0.0 & \mathbf{0.5} & \mathbf{0.5} & 0.0 \\ 0.0 & 0.0 & 0.0 & \mathbf{0.5} & \mathbf{0.5} \\ 0.0 & 0.0 & 0.0 & 0.0 & \mathbf{1.0} \end{bmatrix}$$

HMM3:



Transition matrix

$$\begin{bmatrix} 0.0 & \mathbf{1.0} & 0.0 & 0.0 & 0.0 \\ 0.0 & \mathbf{0.95} & \mathbf{0.05} & 0.0 & 0.0 \\ 0.0 & 0.0 & \mathbf{0.95} & \mathbf{0.05} & 0.0 \\ 0.0 & 0.0 & 0.0 & \mathbf{0.95} & \mathbf{0.05} \\ 0.0 & 0.0 & 0.0 & 0.0 & \mathbf{1.0} \end{bmatrix}$$

HMM model: Θ

In the case of HMMs with Gaussian emission probabilities, the parameter set Θ comprises :

- the transition probabilities a_{ij};
- the parameters of the Gaussian densities characterizing each state, i.e. the means μ_i and the variances Σ_i.

The initial state distribution is sometimes modeled as an additional parameter instead of being represented in the transition matrix.

Size of an HMM model: Ergodic and Gaussian case

In the case of an ergodic HMM with N emitting states and Gaussian emission probabilities, we have :

- $(N-2) \times (N-2)$ transitions, plus $(N-2)$ initial state probabilities and $(N-2)$ probabilities to go to the final state;
- $(N-2)$ emitting states where each pdf is characterized by a D dimensional mean and a $D \times D$ covariance matrix.

Hence, in this case, the total number of parameters is $(N-2) \times (N + D \times (D+1))$.

Outline

1. Introduction
2. Pattern recognition with HMMs
   - Likelihood of a sequence given a HMM
   - Bayesian classification
3. Optimal State Sequence
   - Viterbi Algorithm
4. Acknowledgments

Likelihood of a sequence given a HMM

Likelihood of a sequence given a HMM:

$$p(X|\Theta),$$

i.e. the likelihood of an observation sequence given a model.

Probability of a state sequence

- Assume a state sequence $Q = \{q_1, \cdots, q_T\}$
- *Probability of a state sequence*:

$$P(Q|\Theta) = \prod_{t=1}^{T-1} a_{t,t+1} = a_{1,2} \cdot a_{2,3} \cdots a_{T-1,T}$$

Probability of a state sequence

- Assume a state sequence $Q = \{q_1, \cdots, q_T\}$
- *Probability of a state sequence*:

$$P(Q|\Theta) = \prod_{t=1}^{T-1} a_{t,t+1} = a_{1,2} \cdot a_{2,3} \cdots a_{T-1,T}$$

Likelihood of an observation sequence given a state sequence

- Assume an observation sequence $X = \{x_1, x_2, \cdots, x_T\}$ and a state sequence $Q = \{q_1, \cdots, q_T\}$
- Likelihood of an observation sequence along a single path, Q, for an HMM, Θ:

$$p(X|Q, \Theta) = \prod_{i=1}^{T} p(x_i|q_i, \Theta) = b_1(x_1) \cdot b_2(x_2) \cdots b_T(x_T)$$

Likelihood of an observation sequence given a state sequence

- Assume an observation sequence $X = \{x_1, x_2, \cdots, x_T\}$ and a state sequence $Q = \{q_1, \cdots, q_T\}$
- Likelihood of an observation sequence along a single path, Q, for an HMM, Θ:

$$p(X|Q,\Theta) = \prod_{i=1}^{T} p(x_i|q_i,\Theta) = b_1(x_1) \cdot b_2(x_2) \cdots b_T(x_T)$$

Likelihoods

- *Joint likelihood of an observation sequence X and a path Q*: it consists in the probability that X and Q occur simultaneously, $p(X, Q|\Theta)$, and decomposes into a product of the two quantities defined previously:

$$p(X, Q|\Theta) = p(X|Q, \Theta)P(Q|\Theta) \qquad \text{(Bayes)}$$

- *Likelihood of a sequence with respect to a HMM*: the likelihood of an observation sequence $X = \{x_1, x_2, \cdots, x_T\}$ with respect to a Hidden Markov Model with parameters Θ expands as follows:

$$p(X|\Theta) = \sum_{\text{every possible } Q} p(X, Q|\Theta)$$

Likelihoods

- *Joint likelihood of an observation sequence X and a path Q*: it consists in the probability that X and Q occur simultaneously, $p(X, Q|\Theta)$, and decomposes into a product of the two quantities defined previously:

$$p(X, Q|\Theta) = p(X|Q, \Theta)P(Q|\Theta) \qquad \text{(Bayes)}$$

- *Likelihood of a sequence with respect to a HMM*: the likelihood of an observation sequence $X = \{x_1, x_2, \cdots, x_T\}$ with respect to a Hidden Markov Model with parameters Θ expands as follows:

$$p(X|\Theta) = \sum_{\text{every possible } Q} p(X, Q|\Theta)$$

Forward Recursion

- There is a recursive way to compute $p(X|\Theta)$: Forward Recursion (FR)
- In FR, we define a *forward* variable:

$$p_t(i) = p(x_1, x_2, \cdots x_t, q^t = q_i|\Theta)$$

i.e. $p_t(i)$ is the probability of having observed the partial sequence $\{x_1, x_2, \cdots, x_t\}$ *and* being in the state i at time t, given parameters Θ.

Forward Recursion

- There is a recursive way to compute $p(X|\Theta)$: Forward Recursion (FR)
- In FR, we define a *forward* variable:

$$p_t(i) = p(x_1, x_2, \cdots x_t, q^t = q_i|\Theta)$$

i.e. $p_t(i)$ is the probability of having observed the partial sequence $\{x_1, x_2, \cdots, x_t\}$ *and* being in the state i at time t, given parameters Θ.

Computation of the forward variable

Assume N states with $N-2$ emitting states.

- **Initialization:**
$$p_1(j) = a_{1j} b_j(x_1)$$
with $2 \leq j \leq N-1$
- **Recursion:**
$$p_t(j) = \left[\sum_{i=2}^{N-1} p_{t-1}(i) \cdot a_{ij} \right] b_j(x_t),$$
with $2 \leq t \leq T$ and $2 \leq j \leq N-1$
- **Termination:**
$$p(X|\Theta) = \left[\sum_{i=2}^{N-1} p_T(i) \cdot a_{iN} \right]$$

Computation of the forward variable

Assume N states with $N-2$ emitting states.

- **Initialization:**

$$p_1(j) = a_{1j} b_j(x_1)$$

with $2 \leq j \leq N-1$

- **Recursion:**

$$p_t(j) = \left[\sum_{i=2}^{N-1} p_{t-1}(i) \cdot a_{ij} \right] b_j(x_t),$$

with $2 \leq t \leq T$ and $2 \leq j \leq N-1$

- **Termination:**

$$p(X|\Theta) = \left[\sum_{i=2}^{N-1} p_T(i) \cdot a_{iN} \right]$$

Computation of the forward variable

Assume N states with $N-2$ emitting states.

- **Initialization:**

$$p_1(j) = a_{1j} b_j(x_1)$$

with $2 \leq j \leq N-1$

- **Recursion:**

$$p_t(j) = \left[\sum_{i=2}^{N-1} p_{t-1}(i) \cdot a_{ij} \right] b_j(x_t),$$

with $2 \leq t \leq T$ and $2 \leq j \leq N-1$

- **Termination:**

$$p(X|\Theta) = \left[\sum_{i=2}^{N-1} p_T(i) \cdot a_{iN} \right]$$

Bayesian classification

- Assume that there are many HMMs, Θ_i, $i = 1, \cdots, M$
- Given the likelihood $p(X|\Theta_i)$ computed using the forward recursion algorithm, we can compute the probability of Θ_i, using Bayes' rule:

$$\begin{aligned} P(\Theta_i|X) &= \frac{p(X|\Theta_i)P(\Theta_i)}{P(X|\Theta)} \\ &\propto p(X|\Theta_i)P(\Theta_i) \end{aligned}$$

- Other solution: Maximum likelihood.

Bayesian classification

- Assume that there are many HMMs, Θ_i, $i = 1, \cdots, M$
- Given the likelihood $p(X|\Theta_i)$ computed using the forward recursion algorithm, we can compute the probability of Θ_i, using Bayes' rule:

$$\begin{aligned} P(\Theta_i|X) &= \frac{p(X|\Theta_i)P(\Theta_i)}{P(X|\Theta)} \\ &\propto p(X|\Theta_i)P(\Theta_i) \end{aligned}$$

- Other solution: Maximum likelihood.

Bayesian classification

- Assume that there are many HMMs, Θ_i, $i = 1, \cdots, M$
- Given the likelihood $p(X|\Theta_i)$ computed using the forward recursion algorithm, we can compute the probability of Θ_i, using Bayes' rule:

$$
\begin{aligned}
P(\Theta_i|X) &= \frac{p(X|\Theta_i)P(\Theta_i)}{P(X|\Theta)} \\
&\propto p(X|\Theta_i)P(\Theta_i)
\end{aligned}
$$

- Other solution: Maximum likelihood.

Outline

1. Introduction
2. Pattern recognition with HMMs
   - Likelihood of a sequence given a HMM
   - Bayesian classification
3. Optimal State Sequence
   - Viterbi Algorithm
4. Acknowledgments

Definitions

- *Highest* likelihood $\delta_t(i)$ along a *single* path among all the paths ending in state i at time t:

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} p(q_1, q_2, \cdots, q_{t-1}, q^t = q_i, x_1, x_2, \cdots x_t | \Theta)$$

- *Buffer* $\psi_t(i)$ which allows to keep track of the "best path" ending in state i at time t:

$$\psi_t(i) = \underset{q_1, q_2, \cdots, q_{t-1}}{\operatorname{argmax}} p(q_1, q_2, \cdots, q_{t-1}, q^t = q_i, x_1, x_2, \cdots x_t | \Theta)$$

Definitions

- *Highest* likelihood $\delta_t(i)$ along a *single* path among all the paths ending in state i at time t:

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} p(q_1, q_2, \cdots, q_{t-1}, q^t = q_i, x_1, x_2, \cdots x_t | \Theta)$$

- *Buffer* $\psi_t(i)$ which allows to keep track of the "best path" ending in state i at time t:

$$\psi_t(i) = \operatorname*{argmax}_{q_1, q_2, \cdots, q_{t-1}} p(q_1, q_2, \cdots, q_{t-1}, q^t = q_i, x_1, x_2, \cdots x_t | \Theta)$$

Viterbi Algorithm

1. **Initialization :**

$$
\begin{aligned}
\delta_1(i) &= a_{1i} \cdot b_i(x_1), \quad 2 \leq i \leq N-1 \\
\psi_1(i) &= 0
\end{aligned}
$$

2. **Recursion :**

$$
\begin{aligned}
\delta_{t+1}(j) &= \max_{2 \leq i \leq N-1} [\delta_t(i) \cdot a_{ij}] \cdot b_j(x_{t+1}), \quad \begin{array}{l} 1 \leq t \leq T-1 \\ 2 \leq j \leq N-1 \end{array} \\
\psi_{t+1}(j) &= \arg\max_{2 \leq i \leq N-1} [\delta_t(i) \cdot a_{ij}], \quad \begin{array}{l} 1 \leq t \leq T-1 \\ 2 \leq j \leq N-1 \end{array}
\end{aligned}
$$

3. **Termination :**

$$
\begin{aligned}
p^*(X|\Theta) &= \max_{2 \leq i \leq N-1} [\delta_T(i) \cdot a_{iN}] \\
q_T^* &= \arg\max_{2 \leq i \leq N-1} [\delta_T(i) \cdot a_{iN}]
\end{aligned}
$$

4. **Backtracking :**

$$Q^* = \{q_1^*, \cdots, q_T^*\} \quad \text{so that} \quad q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2,$$

Outline

Acknowledgments

Most of this material is from Hervé Bourlard's course on Speech processing and speech recognition