HMMs for Acoustic Modeling
(Part I)
Lecture 2

Instructor: Preethi Jyothi
Recall: Statistical ASR

Let $O$ be a sequence of acoustic features corresponding to a speech signal. That is, $O = \{O_1, \ldots, O_T\}$, where $O_i \in \mathbb{R}^d$ refers to a d-dimensional acoustic feature vector and $T$ is the length of the sequence.

Let $W$ denote a word sequence. An ASR decoder solves the foll. problem:

$$W^* = \arg \max_W \Pr(W \mid O)$$

$$= \arg \max_W \Pr(O \mid W) \Pr(W)$$
Isolated word recognition

Commonly used acoustic models in ASR systems are Hidden Markov Models (HMMs). Please refer to the comprehensive tutorial of HMMs and their applicability to ASR in the 1980's (with Rabiner, 1989).

Acoustic model:

\[ P(W_{a1}) \]

The HMM is defined by specifying transition probabilities \( a_{ij} \) and observation probabilities \( b_i(o) \), corresponding to the word sequence \( w \) in the form of a sequence of acoustic vectors \( O \).

The most commonly used acoustic models in ASR systems to-date are HMMs. Please refer to Rabiner, 1989.

\[ \text{Compute } \arg \max_w P(O \mid w) \]

Assume the language model is a trigram model.

\[ \text{Pr}(O \mid "up") \]

\[ \text{Pr}(O \mid "down") \]

\[ \text{Pr}(O \mid "left") \]

\[ \text{Pr}(O \mid "right") \]
What are Hidden Markov Models (HMMs)?
Markov Chains

\[ Q = q_1 q_2 \ldots q_N \]  

\[ A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn} \]  

\[ \pi = \pi_1, \pi_2, \ldots, \pi_N \]  

\( A \) is a set of \( N \) states. 

\( A \) is a transition probability matrix \( A \), each \( a_{ij} \) representing the probability of moving from state \( i \) to state \( j \), s.t.  

\[ \sum_{j=1}^{n} a_{ij} = 1 \quad \forall i \]

\( \pi \) is a initial probability distribution over states. \( \pi_i \) is the probability that the Markov chain will start in state \( i \). Some states \( j \) may have \( \pi_j = 0 \), meaning that they cannot be initial states. Also,  

\[ \sum_{i=1}^{n} \pi_i = 1 \]
Ergodic HMM

Bakis network

language processing the lengths of the observations are fixed.

do when they don't have explicit end symbols. This isn't a problem since for most tasks in speech and

numbered state have zero probability). Bakis HMMs are generally used to model

Bakis HMM, no transitions go from a higher-numbered state to a lower-numbered

HMMs, the state transitions proceed from left to right, as shown in Fig.

HMM in Fig.

First, as with a first-order Markov chain, the probability of a particular state depends

given day.

creams Jason ate every day that summer. Our goal is to use these observations to

Imagine that you are a climatologist in the year 2799 studying the history of global

Problem 3 (Learning):

It is also possible to have HMMs without final states or explicit accepting states. Such HMMs define a

Problem 2 in Chapter 8. In the next two sections

used for HMMs doesn't rely on a start or end state, instead representing the distri-

Output Independence:

Second, the probability of an output observation

Notice that in the HMM in Fig. 9.3

A first-order hidden Markov model instantiates two simplifying assumptions.

Markov Assumption:  \( P(q_i|q_1...q_{i-1}) = P(q_i|q_{i-1}) \)

Output Independence:  \( P(o_i|q_1...q_i,...,q_T,o_1,...,o_i,...,o_T) = P(o_i|q_i) \)
## Hidden Markov Model

- **$Q = q_1 q_2 \ldots q_N$** a set of $N$ states
- **$A = a_{11} \ldots a_{ij} \ldots a_{NN}$** a **transition probability matrix** $A$, each $a_{ij}$ representing the probability of moving from state $i$ to state $j$, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \ \forall i$
- **$O = o_1 o_2 \ldots o_T$** a sequence of $T$ **observations**, each one drawn from a vocabulary $V = \{v_1, v_2, \ldots, v_V\}$
- **$B = b_i(o_t)$** a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation $o_t$ being generated from a state $i$
- **$\pi = \pi_1, \pi_2, \ldots, \pi_N$** an **initial probability distribution** over states. $\pi_i$ is the probability that the Markov chain will start in state $i$. Some states $j$ may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$
Three problems for HMMs

| Problem 1 (Likelihood): | Given an HMM $\lambda = (A, B)$ and an observation sequence $O$, determine the likelihood $P(O|\lambda)$. |
|--------------------------|--------------------------------------------------------------------------------------------------|
| Problem 2 (Decoding):    | Given an observation sequence $O$ and an HMM $\lambda = (A, B)$, discover the best hidden state sequence $Q$. |
| Problem 3 (Learning):    | Given an observation sequence $O$ and the set of states in the HMM, learn the HMM parameters $A$ and $B$. |

**Computing Likelihood:** Given an HMM $\lambda = (A, B)$ and an observation sequence $O$, determine the likelihood $P(O|\lambda)$. 

---

“A tutorial on hidden Markov models and selected applications in speech recognition”, Rabiner, 1989
\[
\alpha_t(j) = P(o_1, o_2 \ldots o_t, q_t = j | \lambda) \quad \text{and} \quad \alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t)
\]
9.4 Decoding: The Viterbi Algorithm

For any model, such as an HMM, that contains hidden variables, the task of determining which sequence of variables is the underlying source of some sequence of observations is called the decoding task. In the ice-cream domain, given a sequence of ice-cream observations and an HMM, the task of the decoder is to find the best hidden weather sequence (HHH). More formally,

\[ \text{Decoding: Given as input an HMM } \lambda = (A, B) \text{ and a sequence of observations } O = o_1, o_2, \ldots, o_T, \text{ find the most probable sequence of states } Q = q_1 q_2 q_3 \ldots q_T. \]
Forward Algorithm

1. Initialization:

\[ \alpha_1(j) = \pi_j b_j(o_1) \quad 1 \leq j \leq N \]

2. Recursion:

\[ \alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i)a_{ij}b_j(o_t); \quad 1 \leq j \leq N, 1 < t \leq T \]

3. Termination:

\[ P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \]
Chapter 6 of ice-cream observations

For any model, such as an HMM, that contains hidden variables, the task of determining the probability computation for the current state. Hidden states are in circles, observations in squares. Shaded nodes are included in the observation probability function $t$ for $a$ in the HMM, learn the HMM parameters $A$ and $B$.

Problem 1 (Likelihood): Given an HMM $\lambda = (A, B)$ and an observation sequence $O$, determine the likelihood $P(O|\lambda)$.

Problem 2 (Decoding): Given an observation sequence $O$ and an HMM $\lambda = (A, B)$, discover the best hidden state sequence $Q$.

Problem 3 (Learning): Given an observation sequence $O$ and the set of states in the HMM, learn the HMM parameters $A$ and $B$.

Decoding: Given as input an HMM $\lambda = (A, B)$ and a sequence of observations $O = o_1, o_2, ..., o_T$, find the most probable sequence of states $Q = q_1q_2q_3...q_T$. Three problems for HMMs
Viterbi Trellis

\[ v_t(j) = \max_{q_1, \ldots, q_{t-1}} P(q_1 \ldots q_{t-1}, o_1, o_2 \ldots o_t, q_t = j|\lambda) \]

\[ v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t) \]
Viterbi recursion

1. Initialization:

\[ v_1(j) = \pi_j b_j(o_1) \quad 1 \leq j \leq N \]
\[ b_{t1}(j) = 0 \quad 1 \leq j \leq N \]

2. Recursion

\[ v_t(j) = \max_{i=1}^N v_{t-1}(i) a_{ij} b_j(o_t) \quad 1 \leq j \leq N, 1 < t \leq T \]
\[ b_{t1}(j) = \arg\max_{i=1}^N v_{t-1}(i) a_{ij} b_j(o_t) \quad 1 \leq j \leq N, 1 < t \leq T \]

3. Termination:

The best score:

\[ P^* = \max_{i=1}^N v_T(i) \]

The start of backtrace:

\[ q_{T*} = \arg\max_{i=1}^N v_T(i) \]
Viterbi backtrace

Viterbi backtrace

\[ v_1(2) = 0.32 \]

\[ v_1(1) = 0.02 \]

\[ v_2(2) = \max(0.32 \times 0.12, 0.02 \times 0.10) = 0.038 \]

\[ v_2(1) = \max(0.32 \times 0.20, 0.02 \times 0.25) = 0.064 \]
**Gaussian Observation Model**

- So far, we considered HMMs with discrete outputs
- In acoustic models, HMMs output real valued vectors
- Hence, observation probabilities are defined using probability density functions
- A widely used model: Gaussian distribution

\[ \mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \]

- HMM emission/observation probabilities \( b_j(x) = \mathcal{N}(x | \mu_j, \sigma_j^2) \) where \( \mu_j \) is the mean associated with state \( j \) and \( \sigma_j^2 \) is its variance
- For multivariate Gaussians, \( b_j(x) = \mathcal{N}(x | \mu_j, \Sigma_j) \) where \( \Sigma_j \) is the covariance matrix associated with state \( j \)
Gaussian Mixture Model

• A single Gaussian observation model assumes that the observed acoustic feature vectors are unimodal
Unimodal
Gaussian Mixture Model

- A single Gaussian observation model assumes that the observed acoustic feature vectors are unimodal.
- More generally, we use a “mixture of Gaussians” to model multiple modes in the data.
Mixture Models
Gaussian Mixture Model

- A single Gaussian observation model assumes that the observed acoustic feature vectors are unimodal.

- More generally, we use a “mixture of Gaussians” to model multiple modes in the data.

- Instead of \( b_j(x) = \mathcal{N}(x | \mu_j, \Sigma_j) \) in the single Gaussian case, \( b_j(x) \) now becomes:

\[
b_j(x) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(x | \mu_{jm}, \Sigma_{jm})
\]

where \( c_{jm} \) is the mixing probability for Gaussian component \( m \) of state \( j \),

\[
\sum_{m=1}^{M} c_{jm} = 1, \quad c_{jm} \geq 0
\]