Introduction to Machine Learning

Hidden Markov Models

Andres Mendez-Vazquez

July 8, 2018

Outline

Outline

Markov Random Processes

Markov process (First-Order Markov Model)

A stochastic process is called a Markov process when it has the Markov property:

$$P\left(X_{t_n}|X_{t_{n-1}}=x_{n-1},...X_{t_1}=x_1\right)=P\left(X_{t_n}|X_{t_{n-1}}=x_{n-1}\right) \quad (1)$$

Simply

The future path of a Markov process, given its current state and the past history before, depends only on the current state (not on how this state has been reached).

Thus (Quite an Oversimplification!!!)

A Markov process is characterized by the (one-step) transition probabilities:

$$p_{ij}=P\left(X_n=j|X_{n-1}=i\right) \quad (2)$$

Markov Random Processes

Markov process (First-Order Markov Model)

A stochastic process is called a Markov process when it has the Markov property:

$$P\left(X_{t_n}|X_{t_{n-1}}=x_{n-1},...X_{t_1}=x_1\right)=P\left(X_{t_n}|X_{t_{n-1}}=x_{n-1}\right) \quad (1)$$

Simply

The future path of a Markov process, given its current state and the past history before, depends only on the current state (not on how this state has been reached).

Thus (Quite an Oversimplification!!!)

A Markov process is characterized by the (one-step) transition probabilities:

$$p_{ij}=P\left(X_n=j|X_{n-1}=i\right) \quad (2)$$

Markov Random Processes

Markov process (First-Order Markov Model)

A stochastic process is called a Markov process when it has the Markov property:

$$P\left(X_{t_n}|X_{t_{n-1}}=x_{n-1},...X_{t_1}=x_1\right)=P\left(X_{t_n}|X_{t_{n-1}}=x_{n-1}\right) \quad (1)$$

Simply

The future path of a Markov process, given its current state and the past history before, depends only on the current state (not on how this state has been reached).

Thus (Quite an Oversimplification!!!)

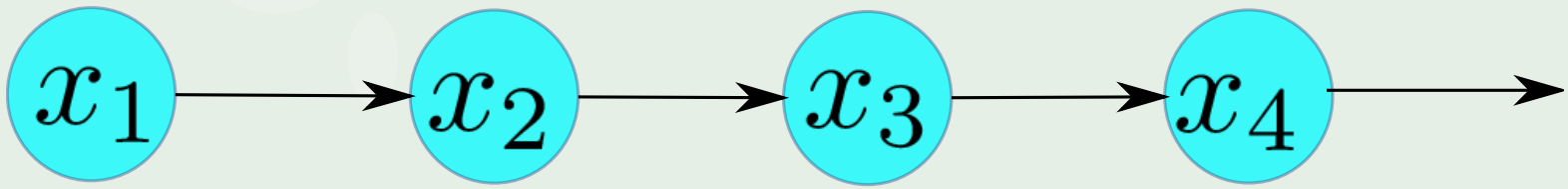
A Markov process is characterized by the (one-step) transition probabilities:

$$p_{ij}=P\left(X_n=j|X_{n-1}=i\right) \quad (2)$$

Graphical Interpretation

A sequence of events depending only in the previous one

$$P\left(x_n|x_{n-1}\right)$$



Markov Chains

Definition (Oversimplified)

A Markov chain is thus a process X_n indexed by integers $n = 0, 1, ...$ such that the states of the system are being indexed by integers $X_n = 0, 1, ...$

From here

The probability of a path $i_1, i_2, ..., i_n$ is

$$P(X_1 = i_1, X_2 = i_2, ..., X_n = i_n) = P(X_1 = i_1) p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{n-1} i_n} \quad (3)$$

Markov Chains

Definition (Oversimplified)

A Markov chain is thus a process X_n indexed by integers $n = 0, 1, ...$ such that the states of the system are being indexed by integers $X_n = 0, 1, ...$

From here

The probability of a path $i_1, i_2, ..., i_n$ is

$$P\left(X_1 = i_1, X_2 = i_2, ..., X_n = i_n\right) = P\left(X_1 = i_1\right) p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{n-1} i_n} \quad (3)$$

Outline

The Transition Probability Matrix of a Markov chain

Thus

The transition probabilities can be arranged as transition probability matrix $\boldsymbol{P} = (p_{i,j})$

$$\text{Final State} \longrightarrow$$

$$\text{Initial State} \downarrow \begin{pmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \boldsymbol{P}$$

Propertities

The row i contains the transition probabilities from state i to other states.

- Since the system always goes to some state, the sum of the column probabilities is 1

Then

A matrix with non-negative elements such that the sum of each column equals 1 is called a stochastic matrix.

The Transition Probability Matrix of a Markov chain

Thus

The transition probabilities can be arranged as transition probability matrix $\boldsymbol{P} = (p_{i,j})$

$$\text{Final State} \longrightarrow$$

$$\text{Initial State} \downarrow \begin{pmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \boldsymbol{P}$$

Propertities

The row i contains the transition probabilities from state i to other states.

- Since the system always goes to some state, the sum of the column probabilities is 1

Then

A matrix with non-negative elements such that the sum of each column equals 1 is called a stochastic matrix.

The Transition Probability Matrix of a Markov chain

Thus

The transition probabilities can be arranged as transition probability matrix $\boldsymbol{P} = (p_{i,j})$

$$\text{Final State} \longrightarrow$$

$$\text{Initial State} \downarrow \begin{pmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \boldsymbol{P}$$

Propertities

The row i contains the transition probabilities from state i to other states.

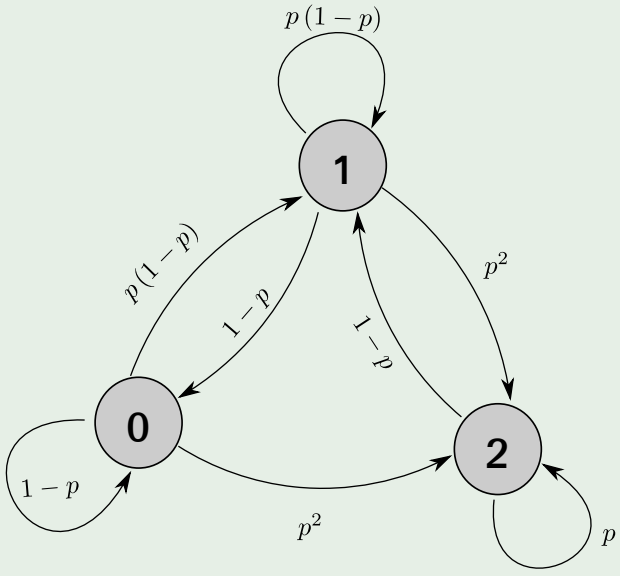
- Since the system always goes to some state, the sum of the column probabilities is 1

Then

A matrix with non-negative elements such that the sum of each column equals 1 is called a stochastic matrix.

Example

We have the following state machine



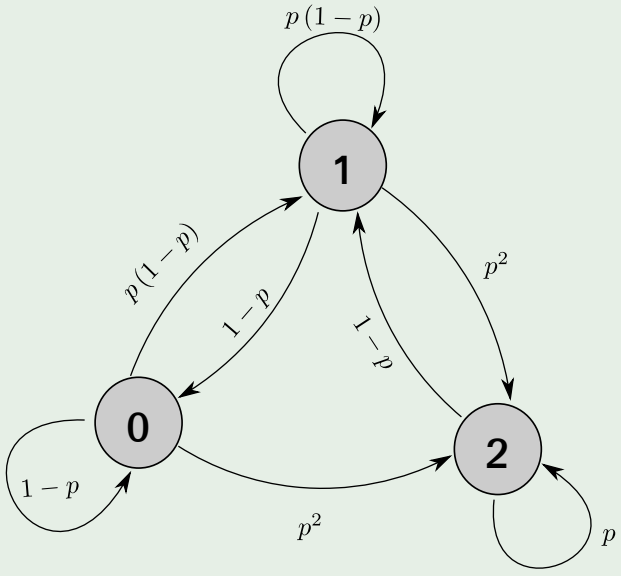
We have the following Stochastic Matrix

$$M = \begin{pmatrix} 1-p & 1-p & 0 \\ p(1-p) & p(1-p) & 1-p \\ p^2 & p^2 & p \end{pmatrix} \tag{4}$$

With $p = \frac{1}{3}$

Example

We have the following state machine



We have the following Stochastic Matrix

$$M = \begin{pmatrix} 1-p & 1-p & 0 \\ p(1-p) & p(1-p) & 1-p \\ p^2 & p^2 & p \end{pmatrix} \tag{4}$$

With $p = \frac{1}{3}$

Using this

We can talk about an initial distribution is given as a row vector π

A stationary probability vector π is defined as a distribution, written as a row vector, that does not change under application of the transition matrix:

$$\pi = P\pi \tag{5}$$

Using this

We can talk about an initial distribution is given as a row vector π

A stationary probability vector π is defined as a distribution, written as a row vector, that does not change under application of the transition matrix:

$$\pi = P\pi \tag{5}$$

Based in the Perron-Frobenious Theorem (Read about it in Wolfran-MathWorld)

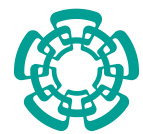
Theorem: Let M be a Markov chain whose left stochastic matrix has no zeros. Then, M has a single stationary distribution, and it converges to this when started at any distribution.

How to find the stationary distribution

We can find this stationary distribution using a slow process called the Power Method

There are others:

- Householder transformations
- Givens rotations
- Arnoldi iteration



How to find the stationary distribution

We can find this stationary distribution using a slow process called the Power Method

There are others:

- Householder transformations
- Givens rotations
- Arnoldi iteration

How to find the stationary distribution

We can find this stationary distribution using a slow process called the Power Method

There are others:

- Householder transformations
- Givens rotations
- Arnoldi iteration

How to find the stationary distribution

We can find this stationary distribution using a slow process called the Power Method

There are others:

- Householder transformations
- Givens rotations
- Arnoldi iteration

How to find the stationary distribution

We can find this stationary distribution using a slow process called the Power Method

There are others:

- Householder transformations
- Givens rotations
- Arnoldi iteration

Outline

Class Sequence and Observations

We have

A sequence of T (feature vectors) observations, $X = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T \rangle$.

In addition

A collection of N classes or hidden states, $\omega_1, \omega_2, ..., \omega_N$

Thus, we have an element like this one

$$\Omega = \langle \omega_1, \omega_2, ..., \omega_T \rangle \tag{6}$$

Note: Total number of class sequences is N^T for sequences of size T.

Class Sequence and Observations

We have

A sequence of T (feature vectors) observations, $X = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T \rangle$.

In addition

A collection of N classes or hidden states, $\omega_1, \omega_2, ..., \omega_N$

Thus, we have an element like this one

$$\Omega = \langle \omega_1, \omega_2, ..., \omega_T \rangle \tag{6}$$

Note: Total number of class sequences is N^T for sequences of size T.

Class Sequence and Observations

We have

A sequence of T (feature vectors) observations, $X = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T \rangle$.

In addition

A collection of N classes or hidden states, $\omega_1, \omega_2, ..., \omega_N$

Thus, we have an element like this one

$$\Omega = \langle \omega_1, \omega_2, ..., \omega_T \rangle \tag{6}$$

Note: Total number of class sequences is N^T for sequences of size T.

Transition Model

Definition of the transition model $\lambda = (A, B, \pi)$ for Hidden Markov Model (HMM)

Where:

1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. $A = \{a_{ij}\}$is the stat transition probability $a_{ij} = P(q_{t+1} = \omega_j | q_t = \omega_i)$ with $1 \leq i, j \leq N$ (In our case $a_{ij} > 0$).
4. $B = \{b_j(x_k)\}$ is the observation symbol probability at state j i.e. $b_j(x_k) = P(x_k | q_t = \omega_j)$ such that $1 \leq j \leq N$ and $1 \leq k \leq M$.
5. The initial state distribution $\pi = \{\pi_i\}$ where $\pi_i = P(q_1 = \omega_i)$ where $1 \leq i \leq N$.

Transition Model

Definition of the transition model $\lambda = (A, B, \pi)$ for Hidden Markov Model (HMM)

Where:

1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. $A = \{a_{ij}\}$is the stat transition probability $a_{ij} = P(q_{t+1} = \omega_j | q_t = \omega_i)$ with $1 \leq i, j \leq N$ (In our case $a_{ij} > 0$).
4. $B = \{b_j(x_k)\}$ is the observation symbol probability at state j i.e. $b_j(x_k) = P(x_k | q_t = \omega_j)$ such that $1 \leq j \leq N$ and $1 \leq k \leq M$.
5. The initial state distribution $\pi = \{\pi_i\}$ where $\pi_i = P(q_1 = \omega_i)$ where $1 \leq i \leq N$.

Transition Model

Definition of the transition model $\lambda = (A, B, \pi)$ for Hidden Markov Model (HMM)

Where:

1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. $A = \{a_{ij}\}$ is the stat transition probability $a_{ij} = P(q_{t+1} = \omega_j | q_t = \omega_i)$ with $1 \leq i, j \leq N$ (In our case $a_{ij} > 0$).
4. $B = \{b_j(x_k)\}$ is the observation symbol probability at state j i.e. $b_j(x_k) = P(x_k | q_t = \omega_j)$ such that $1 \leq j \leq N$ and $1 \leq k \leq M$.
5. The initial state distribution $\pi = \{\pi_i\}$ where $\pi_i = P(q_1 = \omega_i)$ where $1 \leq i \leq N$.

Transition Model

Definition of the transition model $\lambda = (A, B, \pi)$ for Hidden Markov Model (HMM)

Where:

1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. $A = \{a_{ij}\}$is the stat transition probability
   $a_{ij} = P\left(q_{t+1} = \omega_j | q_t = \omega_i\right)$ with $1 \leq i, j \leq N$ (In our case $a_{ij} > 0$).
4. $B = \{b_j\left(x_k\right)\}$ is the observation symbol probability at state j i.e.
   $b_j\left(x_k\right) = P\left(x_k | q_t = \omega_j\right)$ such that $1 \leq j \leq N$ and $1 \leq k \leq M$.
5. The initial state distribution $\pi = \{\pi_i\}$ where $\pi_i = P\left(q_1 = \omega_i\right)$ where $1 \leq i \leq N$.

Transition Model

Definition of the transition model $\lambda = (A, B, \pi)$ for Hidden Markov Model (HMM)

Where:

1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. $A = \{a_{ij}\}$is the stat transition probability
   $a_{ij} = P(q_{t+1} = \omega_j | q_t = \omega_i)$ with $1 \leq i, j \leq N$ (In our case $a_{ij} > 0$).
4. $B = \{b_j(x_k)\}$ is the observation symbol probability at state j i.e.
   $b_j(x_k) = P(x_k | q_t = \omega_j)$ such that $1 \leq j \leq N$ and $1 \leq k \leq M$.
5. The initial state distribution $\pi = \{\pi_i\}$ where $\pi_i = P(q_1 = \omega_i)$ where $1 \leq i \leq N$.

Transition Model

Definition of the transition model $\lambda = (A, B, \pi)$ for Hidden Markov Model (HMM)

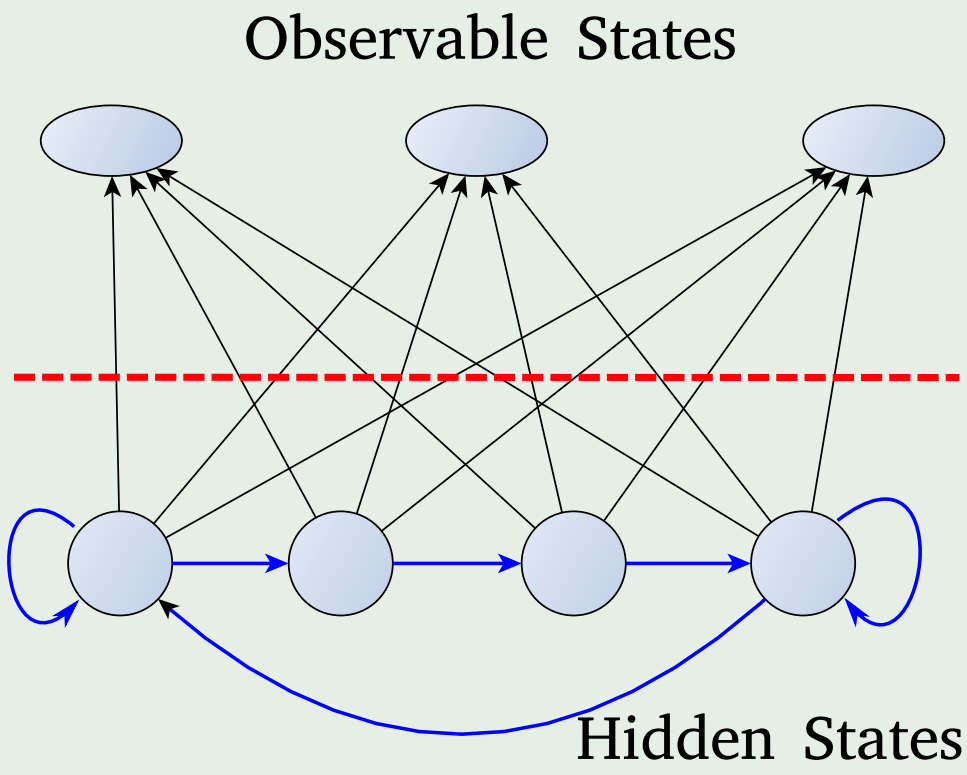
Where:

1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. $A = \{a_{ij}\}$is the stat transition probability $a_{ij} = P(q_{t+1} = \omega_j | q_t = \omega_i)$ with $1 \leq i, j \leq N$ (In our case $a_{ij} > 0$).
4. $B = \{b_j(x_k)\}$ is the observation symbol probability at state j i.e. $b_j(x_k) = P(x_k | q_t = \omega_j)$ such that $1 \leq j \leq N$ and $1 \leq k \leq M$.
5. The initial state distribution $\pi = \{\pi_i\}$ where $\pi_i = P(q_1 = \omega_i)$ where $1 \leq i \leq N$.

Example

We have that

The hidden states conform a probabilistic state machine with emission probability for the symbols.



Outline

Using a HMM

It can be used as a generator of sequences X

1. Choose an intitial state $q_1 = \omega_i$ according to the initial state distribution π.
2. Set $t = 1$.
3. Choose an observation $X_t = x_k$ according to the symbol probability in state ω_i, $b_i(k)$.
4. Transit to a new state $q_{t+1} = \omega_j$ according to the state transition probability distribution for state ω_i, a_{ij}.
5. Set $t = t + 1$, return to step 3 if $t < T$.

Using a HMM

It can be used as a generator of sequences X

1. Choose an intitial state $q_1 = \omega_i$ according to the initial state distribution π.
2. Set $t = 1$.
3. Choose an observation $X_t = x_k$ according to the symbol probability in state ω_i, $b_i(k)$.
4. Transit to a new state $q_{t+1} = \omega_j$ according to the state transition probability distribution for state ω_i, a_{ij}.
5. Set $t = t + 1$, return to step 3 if $t < T$.

Using a HMM

It can be used as a generator of sequences X

1. Choose an intitial state $q_1 = \omega_i$ according to the initial state distribution π.
2. Set $t = 1$.
3. Choose an observation $X_t = x_k$ according to the symbol probability in state ω_i, $b_i(k)$.
4. Transit to a new state $q_{t+1} = \omega_j$ according to the state transition probability distribution for state ω_i, a_{ij}.
5. Set $t = t + 1$, return to step 3 if $t < T$.

Using a HMM

It can be used as a generator of sequences X

1. Choose an intitial state $q_1 = \omega_i$ according to the initial state distribution π.
2. Set $t = 1$.
3. Choose an observation $X_t = x_k$ according to the symbol probability in state ω_i, $b_i(k)$.
4. Transit to a new state $q_{t+1} = \omega_j$ according to the state transition probability distribution for state ω_i, a_{ij}.
5. Set $t = t + 1$, return to step 3 if $t < T$.

Using a HMM

It can be used as a generator of sequences X

1. Choose an intitial state $q_1 = \omega_i$ according to the initial state distribution π.
2. Set $t = 1$.
3. Choose an observation $X_t = x_k$ according to the symbol probability in state ω_i, $b_i(k)$.
4. Transit to a new state $q_{t+1} = \omega_j$ according to the state transition probability distribution for state ω_i, a_{ij}.
5. Set $t = t + 1$, return to step 3 if $t < T$.

How do we implement this

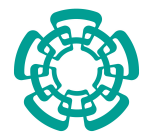
We need to move from q_i to q_{i+1}

- To move from one state to another it is possible to use the following algorithm

Roulette Wheel Selection

1. Allocate an array A of size $n+1$.
2. Set $A[1]=a_{i1}$.
3. For each probability a_{ij} from 1 to $n-1$:

$$A[j] = A[j-1] + p_j \text{ (Cumulative Probability)}$$



How do we implement this

We need to move from q_i to q_{i+1}

- To move from one state to another it is possible to use the following algorithm

Roulette Wheel Selection

1. Allocate an array A of size $n+1$.
2. Set $A[1]=a_{i1}$.
3. For each probability a_{ij} from 1 to $n-1$:

$$A[j] = A[j-1] + p_j \text{ (Cumulative Probability)}$$

Generation

We have

1. Generate a uniformlyrandom value x in the range $[0, 1)$.
2. Using a binary search, find the index j of the smallest element in A larger than x.
3. Return j.

Use the same to emit the symbol

$$b_i\,(k)$$

Generation

We have

1. Generate a uniformlyrandom value x in the range $[0, 1)$.
2. Using a binary search, find the index j of the smallest element in A larger than x.
3. Return j.

Use the same to emit the symbol

$$b_i\left(k\right)$$

Outline

What do we want?

Problem 1

Given the sequence X and a certain model λ, How we efficiently compute $P(X|\lambda)$?

Problem 2

Given the observation sequence X and a model λ, How we choose a corresponding class sequence Ω? Which best explains the observations.

Problem 3

How do we ajust the parameters of the λ to maximize $P(X|\lambda)$?

What do we want?

Problem 1

Given the sequence X and a certain model λ, How we efficiently compute $P(X|\lambda)$?

Problem 2

Given the observation sequence X and a model λ, How we choose a corresponding class sequence Ω? Which best explains the observations.

Problem 3

How do we ajust the parameters of the λ to maximize $P(X|\lambda)$?

What do we want?

Problem 1

Given the sequence X and a certain model λ, How we efficiently compute $P(X|\lambda)$?

Problem 2

Given the observation sequence X and a model λ, How we choose a corresponding class sequence Ω? Which best explains the observations.

Problem 3

How do we ajust the parameters of the λ to maximize $P(X|\lambda)$?

Outline

First Problem

Task

Our task is to decide to which class sequence $\Omega = \langle \omega_1, \omega_2, ..., \omega_T \rangle$ the sequence $X = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T \rangle$ belongs.

For this, first consider a fix sequence of states

$\Omega = \langle \omega_1, \omega_2, ..., \omega_T \rangle$

First Problem

Task

Our task is to decide to which class sequence $\Omega = \langle \omega_1, \omega_2, ..., \omega_T \rangle$ the sequence $X = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T \rangle$ belongs.

For this, first consider a fix sequence of states

$\Omega = \langle \omega_1, \omega_2, ..., \omega_T \rangle$

Outline

First, some assumptions

First

Given the sequence of classes, the observations are statistically independent.

Second

$$P(X|\theta, \lambda) =$$

First, some assumptions

First

Given the sequence of classes, the observations are statistically independent.

Second

The probability density function in one class does not depend on the other classes.

Thus, the probability of the observation sequence X

$$P(X|\theta, \lambda) = P(o_1, o_2, \ldots, o_T|\theta, \lambda)$$

First, some assumptions

First

Given the sequence of classes, the observations are statistically independent.

Second

The probability density function in one class does not depend on the other classes.

Thus, the probability of the observation sequence X

$$P(X|\Omega,\lambda) = P(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T|\omega_1, \omega_2, ..., \omega_T, \lambda)$$

$$= \prod_{k=1}^{T} P(\boldsymbol{x}_k|\omega_1, \omega_2, ..., \omega_T, \lambda)$$

$$= \prod_{k=1}^{T} P(\boldsymbol{x}_k|\omega_k, \lambda)$$

First, some assumptions

First

Given the sequence of classes, the observations are statistically independent.

Second

The probability density function in one class does not depend on the other classes.

Thus, the probability of the observation sequence X

$$\begin{aligned} P(X|\Omega,\lambda) &= P(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T|\omega_1, \omega_2, ..., \omega_T, \lambda) \\ &= \prod_{k=1}^{T} P(\boldsymbol{x}_k|\omega_1, \omega_2, ..., \omega_T, \lambda) \\ &= \prod_{k=1}^{T} P(\boldsymbol{x}_k|\omega_k, \lambda) \end{aligned}$$

First, some assumptions

First

Given the sequence of classes, the observations are statistically independent.

Second

The probability density function in one class does not depend on the other classes.

Thus, the probability of the observation sequence X

$$\begin{aligned} P(X|\Omega,\lambda) =& P(\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_T|\omega_1,\omega_2,...,\omega_T,\lambda) \\ &= \prod_{k=1}^{T} P(\boldsymbol{x}_k|\omega_1,\omega_2,...,\omega_T,\lambda) \\ &= \prod_{k=1}^{T} P(\boldsymbol{x}_k|\omega_k,\lambda) \end{aligned}$$

Then

We need the probability of such a state sequence Ω

$$P(\Omega|\lambda) \tag{7}$$

We can use the Markov Condition

Then

We need the probability of such a state sequence Ω

$$P(\Omega|\lambda) \tag{7}$$

We can use the Markov Condition

$$\begin{aligned} P(\Omega|\lambda) =& P(\omega_1, \omega_2, ..., \omega_T|\lambda) \\ =& P(\omega_T|\omega_1, \omega_2, ..., \omega_{T-1}, \lambda) \times \cdots \\ & P(\omega_{T-1}|\omega_1, \omega_2, ..., \omega_{T-2}, \lambda) \times \cdots \\ & P(\omega_1|\lambda) \end{aligned}$$

Then

We need the probability of such a state sequence Ω

$$P(\Omega|\lambda) \tag{7}$$

We can use the Markov Condition

$$\begin{aligned} P(\Omega|\lambda) =& P(\omega_1, \omega_2, ..., \omega_T|\lambda) \\ =& P(\omega_T|\omega_1, \omega_2, ..., \omega_{T-1}, \lambda) \times \cdots \\ & P(\omega_{T-1}|\omega_1, \omega_2, ..., \omega_{T-2}, \lambda) \times \cdots \\ & P(\omega_1|\lambda) \end{aligned}$$

Thus

$$P(\Omega|\lambda) = P(\omega_1|\lambda) \prod_{k=2}^{N} P(\omega_k|\omega_{k-1}, \lambda) \tag{8}$$

Then

We need the probability of such a state sequence Ω

$$P(\Omega|\lambda) \tag{7}$$

We can use the Markov Condition

$$\begin{aligned} P(\Omega|\lambda) =& P(\omega_1, \omega_2, ..., \omega_T|\lambda) \\ =& P(\omega_T|\omega_1, \omega_2, ..., \omega_{T-1}, \lambda) \times \cdots \\ & P(\omega_{T-1}|\omega_1, \omega_2, ..., \omega_{T-2}, \lambda) \times \cdots \\ & P(\omega_1|\lambda) \end{aligned}$$

Thus

$$P(\Omega|\lambda) = P(\omega_1|\lambda) \prod_{k=2}^{N} P(\omega_k|\omega_{k-1}, \lambda) \tag{8}$$

Using our notation

Then

$$P(\Omega|\lambda) = \pi_1 a_{1,2} \cdots a_{T-1,T} \tag{9}$$

Now

$$P(X|\Omega,\lambda) = b_{q_1}(x_1) \cdots b_{q_T}(x_T) \tag{10}$$

Using our notation

Then

$$P(\Omega|\lambda) = \pi_1 a_{1,2} \cdots a_{T-1,T} \tag{9}$$

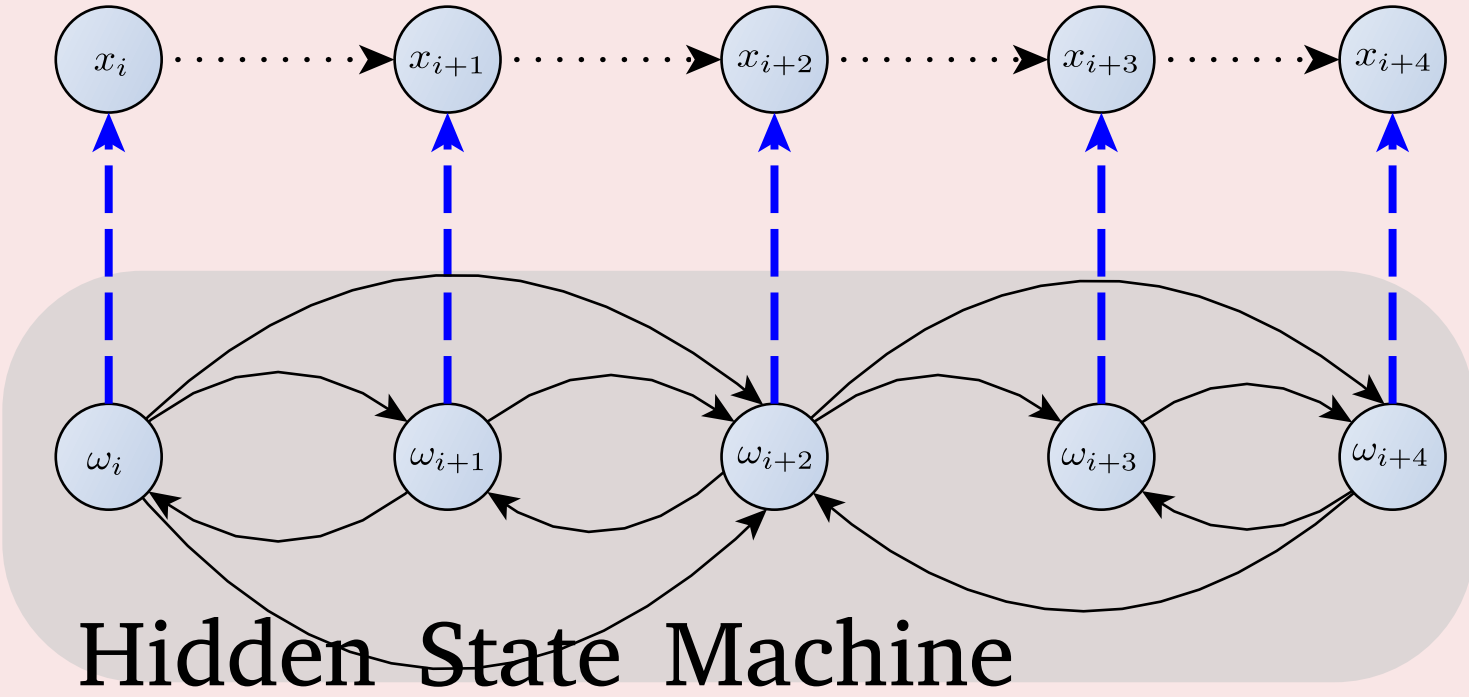
Now

$$P(X|\Omega,\lambda) = b_{q_1}(x_1) \cdots b_{q_T}(x_T) \tag{10}$$

What do we want?

Then, we want to the probability of X and Ω to happen simultaneously

$$P(X, \Omega|\lambda) = P(X|\Omega, \lambda) P(\Omega|\lambda) \tag{11}$$



Outline

Thus, we need to calculate

What?

$$P(X|\lambda) = \sum_{\Omega} P(X, \Omega|\lambda) \tag{12}$$

Using Bayes Rule

$$P(X|\lambda) = \sum_{\Omega} P(X|\Omega, \lambda) P(\Omega|\lambda) \tag{13}$$

Thus, we need to calculate

What?

$$P(X|\lambda) = \sum_{\Omega} P(X, \Omega|\lambda) \tag{12}$$

Using Bayes Rule

$$P(X|\lambda) = \sum_{\Omega} P(X|\Omega, \lambda) P(\Omega|\lambda) \tag{13}$$

Putting All Together

Thus, we have then

$$P(X|\lambda) = \sum_{\Omega} P(X|\Omega, \lambda) P(\Omega|\lambda)$$

$$= \sum_{q_1, q_2, \ldots, q_T} \pi_1 b_{q_1}(x_1) \left[a_{q_1,q_2} b_{q_2}(x_2)\right] \cdots \left[b_{q_T}(x_T)\right]$$

Putting All Together

Thus, we have then

$$P(X|\lambda) = \sum_{\Omega} P(X|\Omega,\lambda)\, P(\Omega|\lambda)$$
$$= \sum_{q_1,q_2,\ldots,q_T} \pi_1 b_{q_1}(x_1)\left[a_{q_1,q_2} b_{q_2}(x_2)\right]\cdots\left[b_{q_T}(x_T)\right]$$

Complexity for brute force approach

Multiplications

We need $2T$ multiplications (Rearranging):

$$\underbrace{\left[\pi_1 a_{q_1,q_2}\cdots a_{q_{T-1},q_T}\right]}_{T}\underbrace{\left[b_{q_1}\left(x_1\right)\cdots b_{q_T}\left(x_T\right)\right]}_{T} \tag{14}$$

Sums

We need N^T additions because of the term $\sum_{q_1,q_2,\cdots,q_T}$.

Total number of Operations

$$O\left(TN^T\right) \tag{15}$$

Complexity for brute force approach

Multiplications

We need $2T$ multiplications (Rearranging):

$$\underbrace{\left[\pi_1 a_{q_1,q_2} \cdots a_{q_{T-1},q_T}\right]}_{T}\underbrace{\left[b_{q_1}\left(x_1\right) \cdots b_{q_T}\left(x_T\right)\right]}_{T} \tag{14}$$

Sums

We need N^T additions because of the term $\sum_{q_1,q_2,\ldots,q_T}$.

Total number of Operations

$$O\left(TN^T\right) \tag{15}$$

Complexity for brute force approach

Multiplications

We need $2T$ multiplications (Rearranging):

$$\underbrace{\left[\pi_1 a_{q_1,q_2} \cdots a_{q_{T-1},q_T}\right]}_{T}\underbrace{\left[b_{q_1}\left(x_1\right) \cdots b_{q_T}\left(x_T\right)\right]}_{T} \tag{14}$$

Sums

We need N^T additions because of the term $\sum_{q_1,q_2,\ldots,q_T}$.

Total number of Operations

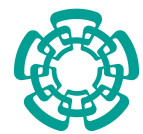
$$O\left(TN^T\right) \tag{15}$$

Outline

This will become part of a Dynamic Programming Solution

Together with a Backward procedure

The forward and backward procedure will help to solve the second problem

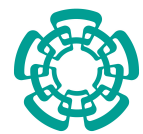


Consider the following

Define

$$\alpha_t(i) = P(x_1, x_2, ..., x_t, q_t = \omega_i | \lambda) \tag{16}$$

Probability of the partial observation until time t and state ω_i.



Consider the following

Define

$$\alpha_t(i) = P(x_1, x_2, ..., x_t, q_t = \omega_i | \lambda) \tag{16}$$

Probability of the partial observation until time t and state ω_i.

We can use an induction, we need to prove

1. Initialization $\alpha_1(i) = \pi_i b_i(x_1)\ 1 \leq i \leq N$
2. Induction $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1})\ 1 \leq t \leq T-1$ and $1 \leq j \leq N$.
3. Termination $P(X|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

Consider the following

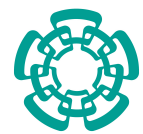
Define

$$\alpha_t(i) = P(x_1, x_2, ..., x_t, q_t = \omega_i | \lambda) \tag{16}$$

Probability of the partial observation until time t and state ω_i.

We can use an induction, we need to prove

1. Initialization $\alpha_1(i) = \pi_i b_i(x_1)\ 1 \leq i \leq N$
2. Induction $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1})\ 1 \leq t \leq T-1$ and $1 \leq j \leq N$.
3. Termination $P(X|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$



Consider the following

Define

$$\alpha_t(i) = P(x_1, x_2, ..., x_t, q_t = \omega_i | \lambda) \tag{16}$$

Probability of the partial observation until time t and state ω_i.

We can use an induction, we need to prove

1. Initialization $\alpha_1(i) = \pi_i b_i(x_1)\ 1 \leq i \leq N$
2. Induction $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1})\ 1 \leq t \leq T-1$ and $1 \leq j \leq N$.
3. Termination $P(X|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

Outline

How do we do this?

Step 1

You need to prove it is correct for $\alpha_1(i)$ with $1 \leq i \leq N$

Assume it is true for t

Then, you need to prove it for $\alpha_{t+1}(j)$ with $1 \leq t \leq T-1$ and $1 \leq j \leq N$.

Then

At termination you have a way to calculate the desired probability $P(X|\lambda)$.

How do we do this?

Step 1

You need to prove it is correct for $\alpha_1(i)$ with $1 \leq i \leq N$

Assume it is true for t

Then, you need to prove it for $\alpha_{t+1}(j)$ with $1 \leq t \leq T-1$ and $1 \leq j \leq N$.

Then

At termination you have a way to calculate the desired probability $P(X|\lambda)$.

How do we do this?

Step 1

You need to prove it is correct for $\alpha_1(i)$ with $1 \leq i \leq N$

Assume it is true for t

Then, you need to prove it for $\alpha_{t+1}(j)$ with $1 \leq t \leq T-1$ and $1 \leq j \leq N$.

Then

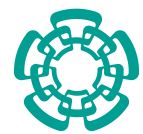
At termination you have a way to calculate the desired probability $P(X|\lambda)$.

Step 1

Initializes the forward probabilities as

$$\alpha_1(i) = P(x_1, q_1 = \omega_i | \lambda) \tag{17}$$

Here is the idea of moving from one state at time t to another state on time $t + 1$

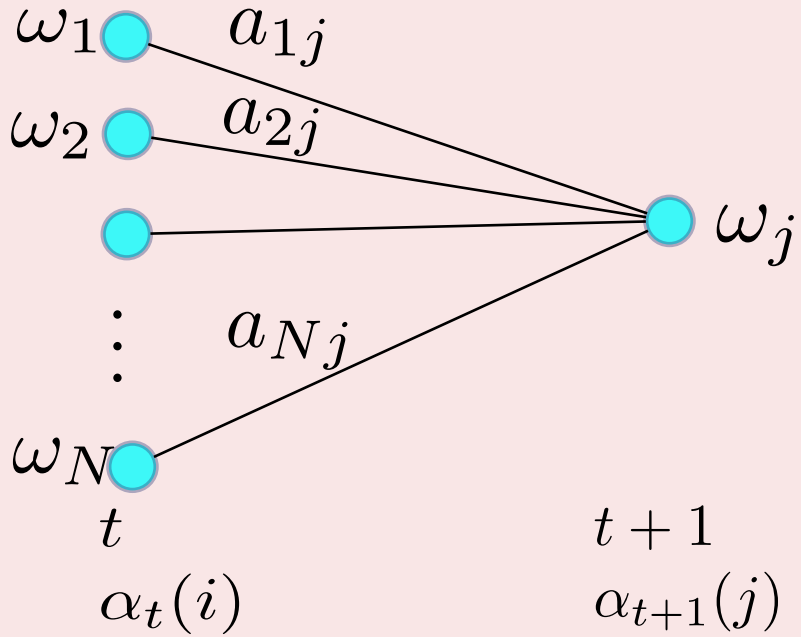


Step 1

Initializes the forward probabilities as

$$\alpha_1(i) = P(x_1, q_1 = \omega_i | \lambda) \tag{17}$$

Here is the idea of moving from one state at time t to another state on time $t+1$

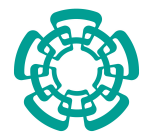


What?

Basically

Our problem is we do not now from which state we start to arrive to j

Now, we start calculate $\alpha_{t+1}(j)$



What?

Basically

Our problem is we do not now from which state we start to arrive to j

Now, we start calculate $\alpha_{t+1}(j)$

$$\alpha_{t+1}(j) = P(x_1, x_2, ..., x_t, x_{t+1}, q_{t+1} = \omega_j | \lambda)$$

$$= P(x_{t+1} | x_1, x_2, ..., x_t, q_{t+1} = \omega_j, \lambda) \times \cdots$$

$$P(x_1, x_2, ..., x_t, q_{t+1} = \omega_j | \lambda)$$

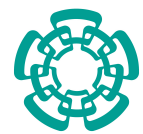
What?

Basically

Our problem is we do not now from which state we start to arrive to j

Now, we start calculate $\alpha_{t+1}(j)$

$$\begin{aligned}\alpha_{t+1}(j) =& P(x_1, x_2, ..., x_t, x_{t+1}, q_{t+1} = \omega_j|\lambda) \\ =& P(x_{t+1}|x_1, x_2, ..., x_t, q_{t+1} = \omega_j, \lambda) \times \cdots \\ & P(x_1, x_2, ..., x_t, q_{t+1} = \omega_j|\lambda)\end{aligned}$$



We need to calculate

First

$$P\left(x_{t+1}|x_1, x_2, ..., x_t, q_{t+1} = \omega_j, \lambda\right) \tag{18}$$

Second

$$P\left(x_1, x_2, ..., x_t, q_{t+1} = \omega_j|\lambda\right) \tag{19}$$

We need to calculate

First

$$P\left(x_{t+1}|x_1, x_2, ..., x_t, q_{t+1} = \omega_j, \lambda\right) \tag{18}$$

Second

$$P\left(x_1, x_2, ..., x_t, q_{t+1} = \omega_j|\lambda\right) \tag{19}$$

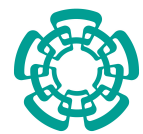
Thus, we start with $P(x_1, x_2, ..., x_t, q_{t+1} = \omega_j|\lambda)$

We have that

$$P(x_1, x_2, ..., x_t, q_{t+1} = \omega_j|\lambda) =$$

$$= \sum_{i=1}^{N} P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda)$$

The second step is due to applying the marginal formula over all possible states



Thus, we start with $P(x_1, x_2, ..., x_t, q_{t+1} = \omega_j|\lambda)$

We have that

$$P(x_1, x_2, ..., x_t, q_{t+1} = \omega_j|\lambda) =$$

$$= \sum_{i=1}^{N} P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda)$$

The second step is due to applying the marginal formula over all possible states

Now, $P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda)$

Thus

$$P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda) =$$

$$= P(q_{t+1} = \omega_j|x_1, x_2, ..., x_t, q_t = \omega_i, \lambda) \times \cdots$$
$$P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda)$$
$$= P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda) \times \cdots$$
$$P(q_{t+1} = \omega_j|q_t = \omega_i, \lambda)$$
$$= \alpha_t(i)\, a_{ij}$$

Now, $P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda)$

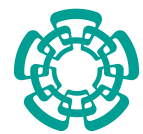
Thus

$$P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda) =$$
$$= P(q_{t+1} = \omega_j|x_1, x_2, ..., x_t, q_t = \omega_i, \lambda) \times \cdots$$
$$P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda)$$
$$= P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda) \times \cdots$$
$$P(q_{t+1} = \omega_j|q_t = \omega_i, \lambda)$$
$$= \alpha_t(i)\, a_{ij}$$

Now, $P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda)$

Thus

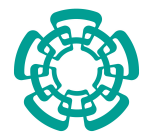
$$
\begin{aligned}
P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda) = & \\
&= P(q_{t+1} = \omega_j|x_1, x_2, ..., x_t, q_t = \omega_i, \lambda) \times \cdots \\
&\quad P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda) \\
&= P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda) \times \cdots \\
&\quad P(q_{t+1} = \omega_j|q_t = \omega_i, \lambda) \\
&= \alpha_t(i)\, a_{ij}
\end{aligned}
$$



Now, $P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda)$

Thus

$$
\begin{aligned}
P(x_1, x_2, ..., x_t, q_t = \omega_i, q_{t+1} = \omega_j|\lambda) = & \\
&= P(q_{t+1} = \omega_j|x_1, x_2, ..., x_t, q_t = \omega_i, \lambda) \times \cdots \\
&\quad P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda) \\
&= P(x_1, x_2, ..., x_t, q_t = \omega_i|\lambda) \times \cdots \\
&\quad P(q_{t+1} = \omega_j|q_t = \omega_i, \lambda) \\
&= \alpha_t(i)\, a_{ij}
\end{aligned}
$$



Thus

We have that

$$P\left(x_1, x_2, ..., x_t, q_{t+1} = \omega_j | \lambda\right) = \sum_{i=1}^{N} \alpha_t\left(i\right) a_{ij} \tag{20}$$

What about the term

$$P\left(x_{t+1} | x_1, x_2, ..., x_t, q_{t+1} = \omega_j, \lambda\right) = P\left(x_{t+1} | q_{t+1} = \omega_j, \lambda\right) = b_j\left(x_{t+1}\right) \tag{21}$$

Putting all together

$$\alpha_{t+1}\left(j\right) = \left[\sum_{i=1}^{N} \alpha_t\left(i\right) a_{ij}\right] b_j\left(x_{t+1}\right) \tag{22}$$

Thus

We have that

$$P\left(x_1, x_2, ..., x_t, q_{t+1} = \omega_j | \lambda\right) = \sum_{i=1}^{N} \alpha_t\left(i\right) a_{ij} \tag{20}$$

What about the term

$$P\left(x_{t+1} | x_1, x_2, ..., x_t, q_{t+1} = \omega_j, \lambda\right) = P\left(x_{t+1} | q_{t+1} = \omega_j, \lambda\right) = b_j\left(x_{t+1}\right) \tag{21}$$

Putting all together

$$\alpha_{t+1}\left(j\right) = \left[\sum_{i=1}^{N} \alpha_t\left(i\right) a_{ij}\right] b_j\left(x_{t+1}\right) \tag{22}$$

Thus

We have that

$$P\left(x_1, x_2, ..., x_t, q_{t+1} = \omega_j | \lambda\right) = \sum_{i=1}^{N} \alpha_t\left(i\right) a_{ij} \tag{20}$$

What about the term

$$P\left(x_{t+1} | x_1, x_2, ..., x_t, q_{t+1} = \omega_j, \lambda\right) = P\left(x_{t+1} | q_{t+1} = \omega_j, \lambda\right) = b_j\left(x_{t+1}\right) \tag{21}$$

Putting all together

$$\alpha_{t+1}\left(j\right) = \left[\sum_{i=1}^{N} \alpha_t\left(i\right) a_{ij}\right] b_j\left(x_{t+1}\right) \tag{22}$$

Finally

This true for

$t = 1, 2, ..., T-1$ and $1 \leq j \leq N$.

Now, applying the marginal idea for $P(x_1, x_2, ..., x_T|\lambda)$

$$P(x_1, x_2, ..., x_T|\lambda) = \sum_{i=1}^{N} P(x_1, x_2, ..., x_T, q_T = \omega_i|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (23)$$

Finally

This true for

$t = 1, 2, ..., T-1$ and $1 \leq j \leq N$.

Now, applying the marginal idea for $P(x_1, x_2, ..., x_T|\lambda)$

$$P(x_1, x_2, ..., x_T|\lambda) = \sum_{i=1}^{N} P(x_1, x_2, ..., x_T, q_T = \omega_i|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{23}$$

Complexity

Given that we are requiered to compute the $\alpha_t(j)$

For $1 \leq t \leq T$ and $1 \leq j \leq N$

We can calculate this using a table (Dynamic Programming Idea)

t \ j	1	2	$\cdots$	N
1	$\alpha_1(1)$	$\alpha_1(2)$	$\cdots$	$\alpha_1(N)$
2	$\alpha_2(1)$	$\underbrace{\alpha_2(1)}_{\text{Cost } N}$	$\cdots$	$\alpha_2(N)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T	$\alpha_T(1)$	$\alpha_T(2)$	$\cdots$	$\alpha_T(N)$

Complexity

Given that we are requiered to compute the $\alpha_t(j)$

For $1 \leq t \leq T$ and $1 \leq j \leq N$

We can calculate this using a table (Dynamic Programming Idea)

t \ i	1	2	$\cdots$	N
1	$\alpha_1(1)$	$\alpha_1(2)$	$\cdots$	$\alpha_1(N)$
2	$\alpha_2(1)$	$\underbrace{\alpha_2(1)}_{\text{Cost } N}$	$\cdots$	$\alpha_2(N)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T	$\alpha_T(1)$	$\alpha_T(2)$	$\cdots$	$\alpha_T(N)$

Complexity

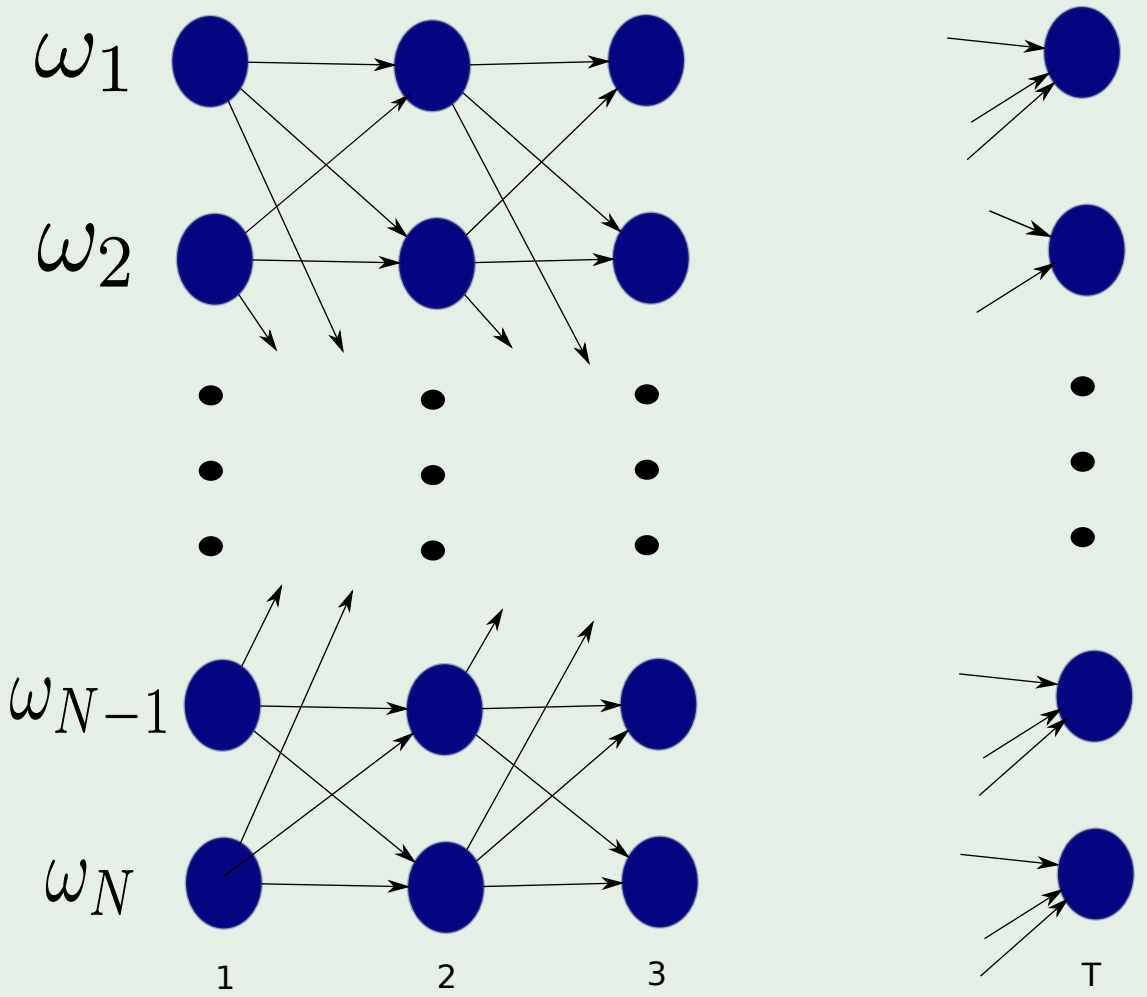
We have

1. Inner loop $O(N)$
2. Outer loop $O(NT)$
   - Then $O(N^2T)$

Outline

All this come from the Lattice Structure

Lattice with Lenght T over time and Depth N over the structure of the probabilistic state machine



Notice the following

The forward method

It is enough to solve the first problem.

However for the second problem

We need to define the backward variable,

$$\beta_t(i) = P(x_{t+1}, x_{t+2}, \ldots, x_T | q_t = \omega_i, \lambda).$$

Notice the following

The forward method

It is enough to solve the first problem.

However for the second problem

We need to define the backward variable,
$\beta_t(i) = P(x_{t+1}, x_{t+2}, ..., x_T | q_t = \omega_i, \lambda)$.

This can be solved in a similar way

Initialization

$$\beta_T(i) = 1, \ 1 \leq i \leq N \tag{24}$$

Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_{t+1}(j) \tag{25}$$

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

This can be solved in a similar way

Initialization

$$\beta_T(i) = 1, \ 1 \leq i \leq N \tag{24}$$

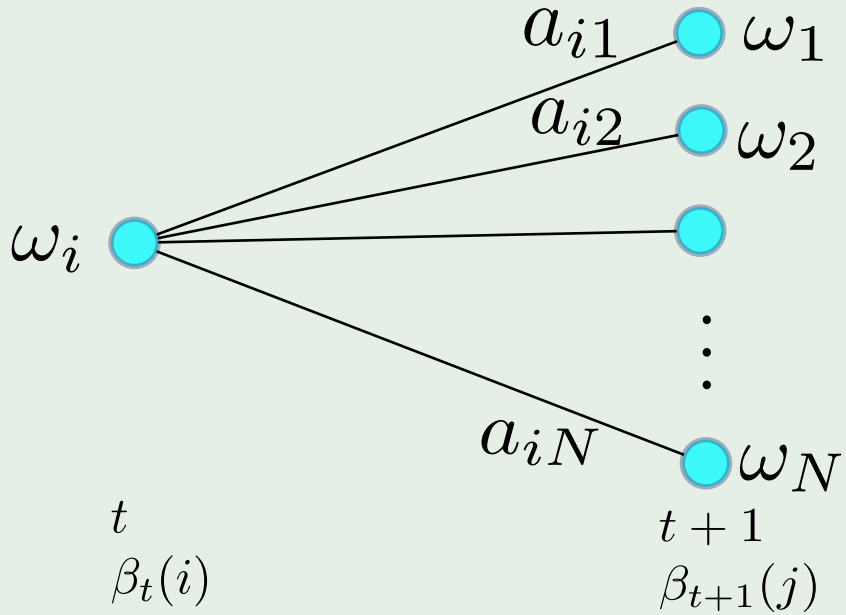
Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_{t+1}(j) \tag{25}$$

For $t = T-1, T-2, ..., 1$ with $1 \leq i \leq N$

Example of the backward process

Example



How?

That is for the next midterm

Please try it for yourselves!!!

Outline

Here, we have several ways of solving the problem

Problem

Given the observation sequence X and a model λ, How we choose a corresponding class sequence Ω?

For this, we define

$$\gamma_t(i) = P(q_t = \omega_i | X, \lambda) \tag{26}$$

Meaning

The probability of being in state ω_i at time t given the observation sequence X and the model λ.

Here, we have several ways of solving the problem

Problem

Given the observation sequence X and a model λ, How we choose a corresponding class sequence Ω?

For this, we define

$$\gamma_t(i) = P(q_t = \omega_i | X, \lambda) \tag{26}$$

Meaning

The probability of being in state ω_i at time t given the observation sequence X and the model λ.

Here, we have several ways of solving the problem

Problem

Given the observation sequence X and a model λ, How we choose a corresponding class sequence Ω?

For this, we define

$$\gamma_t(i) = P(q_t = \omega_i | X, \lambda) \tag{26}$$

Meaning

The probability of being in state ω_i at time t given the observation sequence X and the model λ.

We can find its meaning using Bayes

We get

$$\gamma_t(i) = \frac{P(x_1, ..., x_t, ..., x_T, q_t = \omega_i|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)}$$

$$= \frac{P(x_1, ..., x_t, q_t = \omega_i, x_{t+1} ..., x_T|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)}$$

$$= \frac{P(x_1, ..., x_t, x_{t+1} ..., x_T|q_t = \omega_i, \lambda) P(q_t = \omega_i|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)}$$

$$= \frac{P(x_1, ..., x_t|q_t = \omega_i, \lambda) P(x_{t+1} ..., x_T|q_t = \omega_i, \lambda) P(q_t = \omega_i|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)}$$

$$= \frac{P(x_1, ..., x_t, q_t = \omega_i|\lambda) P(x_{t+1} ..., x_T|q_t = \omega_i, \lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)}$$

$$= \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^{N} P(x_1, ..., x_t, ..., x_T, q_t = \omega_j|\lambda)}$$

$$= \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j) \beta_t(j)}$$

We can find its meaning using Bayes

We get

$$\begin{aligned}\gamma_t(i) &= \frac{P(x_1,...,x_t,...,x_T,q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\ &= \frac{P(x_1,...,x_t,q_t=\omega_i,x_{t+1}...,x_T|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\ &= \frac{P(x_1,...,x_t,x_{t+1}...,x_T|q_t=\omega_i,\lambda)P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\ &= \frac{P(x_1,...,x_t|q_t=\omega_i,\lambda)P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\ &= \frac{P(x_1,...,x_t,q_t=\omega_i|\lambda)P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}P(x_1,...,x_t,...,x_T,q_t=\omega_j|\lambda)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}\end{aligned}$$

We can find its meaning using Bayes

We get

$$\begin{aligned}
\gamma_t(i) &= \frac{P(x_1, ..., x_t, ..., x_T, q_t = \omega_i|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)} \\
&= \frac{P(x_1, ..., x_t, q_t = \omega_i, x_{t+1}..., x_T|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)} \\
&= \frac{P(x_1, ..., x_t, x_{t+1}..., x_T|q_t = \omega_i, \lambda) P(q_t = \omega_i|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)} \\
&= \frac{P(x_1, ..., x_t|q_t = \omega_i, \lambda) P(x_{t+1}..., x_T|q_t = \omega_i, \lambda) P(q_t = \omega_i|\lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)} \\
&= \frac{P(x_1, ..., x_t, q_t = \omega_i|\lambda) P(x_{t+1}..., x_T|q_t = \omega_i, \lambda)}{P(x_1, ..., x_t, ..., x_T|\lambda)} \\
&= \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^{N} P(x_1, ..., x_t, ..., x_T, q_t = \omega_j|\lambda)} \\
&= \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j) \beta_t(j)}
\end{aligned}$$

We can find its meaning using Bayes

We get

$$\begin{aligned}
\gamma_t(i) &= \frac{P(x_1,...,x_t,...,x_T,q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,q_t=\omega_i,x_{t+1}...,x_T|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,x_{t+1}...,x_T|q_t=\omega_i,\lambda)P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t|q_t=\omega_i,\lambda)P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)}
\end{aligned}$$

$$\begin{aligned}
&= \frac{P(x_1,...,x_t,q_t=\omega_i|\lambda)P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}P(x_1,...,x_t,...,x_T,q_t=\omega_j|\lambda)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}
\end{aligned}$$

We can find its meaning using Bayes

We get

$$\begin{aligned}
\gamma_t(i) &= \frac{P(x_1,...,x_t,...,x_T,q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,q_t=\omega_i,x_{t+1}...,x_T|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,x_{t+1}...,x_T|q_t=\omega_i,\lambda)\,P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t|q_t=\omega_i,\lambda)\,P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)\,P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,q_t=\omega_i|\lambda)\,P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N}P(x_1,...,x_t,...,x_T,q_t=\omega_j|\lambda)} \\
&= \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\,\beta_t(j)}
\end{aligned}$$

We can find its meaning using Bayes

We get

$$\begin{aligned}
\gamma_t(i) &= \frac{P(x_1,...,x_t,...,x_T,q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,q_t=\omega_i,x_{t+1}...,x_T|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,x_{t+1}...,x_T|q_t=\omega_i,\lambda)\,P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t|q_t=\omega_i,\lambda)\,P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)\,P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,q_t=\omega_i|\lambda)\,P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N}P(x_1,...,x_t,...,x_T,q_t=\omega_j|\lambda)} \\
&= \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\,\beta_t(j)}
\end{aligned}$$

We can find its meaning using Bayes

We get

$$\begin{aligned}
\gamma_t(i) &= \frac{P(x_1,...,x_t,...,x_T,q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,q_t=\omega_i,x_{t+1}...,x_T|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,x_{t+1}...,x_T|q_t=\omega_i,\lambda)\,P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t|q_t=\omega_i,\lambda)\,P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)\,P(q_t=\omega_i|\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{P(x_1,...,x_t,q_t=\omega_i|\lambda)\,P(x_{t+1}...,x_T|q_t=\omega_i,\lambda)}{P(x_1,...,x_t,...,x_T|\lambda)} \\
&= \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N}P(x_1,...,x_t,...,x_T,q_t=\omega_j|\lambda)} \\
&= \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\,\beta_t(j)}
\end{aligned}$$

In addition, we have that

The $\gamma_t(i)$ are a probability because

$$\sum_{i=1}^{N} \gamma_t(i) = 1 \tag{27}$$

Thus, we can use this to generate the most likely state q_t at time t

$$q_t = \text{argmax}_{1 \leq i \leq N} [\gamma_t(i)], \text{ for } 1 \leq t \leq T \tag{28}$$

In addition, we have that

The $\gamma_t(i)$ are a probability because

$$\sum_{i=1}^{N} \gamma_t(i) = 1 \tag{27}$$

Thus, we can use this to generate the most likely state q_t at time t

$$q_t = \text{argmax}_{1 \leq i \leq N} \left[\gamma_t(i)\right], \text{ for } 1 \leq t \leq T \tag{28}$$

Outline

What can we use? Dynamic Programming

Optimal Substructure

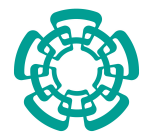
Optimal solution to sub-problems.

Overlapping Sub-problems

Use previous solutions to reduce complexity.

Finally

Reconstructing an optimal solution.



What can we use? Dynamic Programming

Optimal Substructure

Optimal solution to sub-problems.

Overlapping Sub-problems

Use previous solutions to reduce complexity.

Finally

Reconstructing an optimal solution.

What can we use? Dynamic Programming

Optimal Substructure

Optimal solution to sub-problems.

Overlapping Sub-problems

Use previous solutions to reduce complexity.

Finally

Reconstructing an optimal solution.

Optimal Substructure Process

Process

- Show that a solution to the problem consists of making a choice. It leaves a sub-problem to solve.
- Given a problem you suppose an optimal solution.
- Given this choice you determine how to solve the sub-problems.
- Then, you determine that the solution to the sub-problem is optimal by contradiction.

Optimal Substructure Process

Process

- Show that a solution to the problem consists of making a choice. It leaves a sub-problem to solve.
- Given a problem you suppose an optimal solution.
- Given this choice you determine how to solve the sub-problems.
- Then, you determine that the solution to the sub-problem is optimal by contradiction.

Optimal Substructure Process

Process

- Show that a solution to the problem consists of making a choice. It leaves a sub-problem to solve.
- Given a problem you suppose an optimal solution.
- Given this choice you determine how to solve the sub-problems.
- Then, you determine that the solution to the sub-problem is optimal by contradiction.

Optimal Substructure Process

Process

- Show that a solution to the problem consists of making a choice. It leaves a sub-problem to solve.
- Given a problem you suppose an optimal solution.
- Given this choice you determine how to solve the sub-problems.
- Then, you determine that the solution to the sub-problem is optimal by contradiction.

Possible Solutions

One Optimal Criteria

It is possible to to solve for the state sequences that maximizes the expected number of correct pairs of states (q_t, q_{t+1})

However

The most widely criteria is to find the single best path to maximize $P(\Omega|X,\lambda) \approx P(\Omega,X|\lambda)$

Possible Solutions

One Optimal Criteria

It is possible to to solve for the state sequences that maximizes the expected number of correct pairs of states (q_t, q_{t+1})

However

The most widely criteria is to find the single best path to maximize $P(\Omega|X,\lambda) \approx P(\Omega, X|\lambda)$

Outline

Viterbi Algorithm

Setup

We want the best single state sequence $\Omega_i = \{\omega_{i_1}, \omega_{i_2}, ..., \omega_{i_T}\}$ for a given observation X

Thus, we define

$$\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t | \lambda] \quad (29)$$

Viterbi Algorithm

Setup

We want the best single state sequence $\Omega_i = \{\omega_{i_1}, \omega_{i_2}, ..., \omega_{i_T}\}$ for a given observation X

Thus, we define

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t | \lambda] \quad (29)$$

Use Induction

We can do the following

$$\delta_{t+1}(j) = \max_{q_1, q_2, \ldots, q_t} P[q_1, q_2, \cdots, q_{t+1} = j, x_1, x_2, \cdots, x_{t+1} | \lambda]$$

$$= \max_i \{P[q_{t+1} = j, x_{t+1} | q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t, \lambda] \times \cdots$$

$$P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t | \lambda]\}$$

$$= \max_i \{P[q_{t+1} = j, x_{t+1} | q_t = i, \lambda] \times \cdots$$

$$P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t | \lambda]\}$$

$$= \max_i \{P[q_{t+1} = j | q_t = i, \lambda] P[x_{t+1} | q_{t+1} = j, q_t = i, \lambda] \delta_t(i)\}$$

$$= \max_i \{a_{ij} P[x_{t+1} | q_{t+1} = j, \lambda] \delta_t(i)\}$$

$$= \max_i \{a_{ij} b_j(x_{t+1}) \delta_t(i)\} = \max_i \{a_{ij} \delta_t(i)\} b_j(x_{t+1})$$

Use Induction

We can do the following

$$\begin{aligned}\delta_{t+1}(j) &= \max_{q_1,q_2,\cdots,q_t} P[q_1, q_2, \cdots, q_{t+1} = j, x_1, x_2, \cdots, x_{t+1}|\lambda] \\ &= \max_i \{P[q_{t+1} = j, x_{t+1}|q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t, \lambda] \times \cdots \\ &\quad P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda]\} \\ &= \max_i \{P[q_{t+1} = j, x_{t+1}|q_t = i, \lambda] \times \cdots \\ &\quad P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda]\} \\ &= \max_i \{P[q_{t+1} = j|q_t = i, \lambda] P[x_{t+1}|q_{t+1} = j, q_t = i, \lambda] \delta_t(i)\} \\ &= \max_i \{a_{ij} P[x_{t+1}|q_{t+1} = j, \lambda] \delta_t(i)\} \\ &= \max_i \{a_{ij} b_j(x_{t+1}) \delta_t(i)\} = \max_i \{a_{ij} \delta_t(i)\} b_j(x_{t+1})\end{aligned}$$

Use Induction

We can do the following

$$\begin{aligned}
\delta_{t+1}(j) &= \max_{q_1,q_2,\cdots,q_t} P[q_1, q_2, \cdots, q_{t+1} = j, x_1, x_2, \cdots, x_{t+1}|\lambda] \\
&= \max_i \{P[q_{t+1} = j, x_{t+1}|q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t, \lambda] \times \cdots \\
&\quad P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda]\} \\
&= \max_i \{P[q_{t+1} = j, x_{t+1}|q_t = i, \lambda] \times \cdots \\
&\quad P[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda]\} \\
&= \max_i \{P[q_{t+1} = j|q_t = i, \lambda] P[x_{t+1}|q_{t+1} = j, q_t = i, \lambda] \delta_t(i)\} \\
&= \max_i \{a_{ij} P[x_{t+1}|q_{t+1} = j, \lambda] \delta_t(i)\} \\
&= \max_i \{a_{ij} b_j(x_{t+1}) \delta_t(i)\} = \max_i \{a_{ij} \delta_t(i)\} b_j(x_{t+1})
\end{aligned}$$

Use Induction

We can do the following

$$\begin{aligned}
\delta_{t+1}(j) &= \max_{q_1,q_2,\ldots,q_t} P\left[q_1, q_2, \cdots, q_{t+1} = j, x_1, x_2, \cdots, x_{t+1}|\lambda\right] \\
&= \max_i \{P\left[q_{t+1} = j, x_{t+1}|q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t, \lambda\right] \times \cdots \\
&\quad P\left[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda\right]\} \\
&= \max_i \{P\left[q_{t+1} = j, x_{t+1}|q_t = i, \lambda\right] \times \cdots \\
&\quad P\left[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda\right]\} \\
&= \max_i \{P\left[q_{t+1} = j|q_t = i, \lambda\right] P\left[x_{t+1}|q_{t+1} = j, q_t = i, \lambda\right] \delta_t(i)\} \\
&= \max_i \{a_{ij} P\left[x_{t+1}|q_{t+1} = j, \lambda\right] \delta_t(i)\} \\
&= \max_i \{a_{ij} b_j(x_{t+1}) \delta_t(i)\} = \max_i \{a_{ij} \delta_t(i)\} b_j(x_{t+1})
\end{aligned}$$

Use Induction

We can do the following

$$\begin{aligned}
\delta_{t+1}(j) &= \max_{q_1,q_2,\ldots,q_t} P\left[q_1, q_2, \cdots, q_{t+1} = j, x_1, x_2, \cdots, x_{t+1}|\lambda\right] \\
&= \max_i \{P\left[q_{t+1} = j, x_{t+1}|q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t, \lambda\right] \times \cdots \\
&\quad P\left[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda\right]\} \\
&= \max_i \{P\left[q_{t+1} = j, x_{t+1}|q_t = i, \lambda\right] \times \cdots \\
&\quad P\left[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda\right]\} \\
&= \max_i \{P\left[q_{t+1} = j|q_t = i, \lambda\right] P\left[x_{t+1}|q_{t+1} = j, q_t = i, \lambda\right] \delta_t(i)\} \\
&= \max_i \{a_{ij} P\left[x_{t+1}|q_{t+1} = j, \lambda\right] \delta_t(i)\} \\
&= \max_i \{a_{ij} b_j(x_{t+1}) \delta_t(i)\} = \max_i \{a_{ij} \delta_t(i)\} b_j(x_{t+1})
\end{aligned}$$

Use Induction

We can do the following

$$\begin{aligned}
\delta_{t+1}(j) &= \max_{q_1,q_2,\ldots,q_t} P\left[q_1, q_2, \cdots, q_{t+1} = j, x_1, x_2, \cdots, x_{t+1}|\lambda\right] \\
&= \max_i \{P\left[q_{t+1} = j, x_{t+1}|q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t, \lambda\right] \times \cdots \\
&\quad P\left[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda\right]\} \\
&= \max_i \{P\left[q_{t+1} = j, x_{t+1}|q_t = i, \lambda\right] \times \cdots \\
&\quad P\left[q_1, q_2, \cdots, q_t = i, x_1, x_2, \cdots, x_t|\lambda\right]\} \\
&= \max_i \{P\left[q_{t+1} = j|q_t = i, \lambda\right] P\left[x_{t+1}|q_{t+1} = j, q_t = i, \lambda\right] \delta_t(i)\} \\
&= \max_i \{a_{ij} P\left[x_{t+1}|q_{t+1} = j, \lambda\right] \delta_t(i)\} \\
&= \max_i \{a_{ij} b_j(x_{t+1}) \delta_t(i)\} = \max_i \{a_{ij}\delta_t(i)\} b_j(x_{t+1})
\end{aligned}$$

Outline

Procedure

Initialization

- $\delta_1(i) = \pi_i b_i(x_1),\ 1 \leq i \leq N$
- Array $\Psi_1(i) = 0,\ 1 \leq i \leq N$

Procedure

Initialization

- $\delta_1(i) = \pi_i b_i(x_1)$, $1 \leq i \leq N$
- Array $\Psi_1(i) = 0$, $1 \leq i \leq N$

Recusrions

- $\delta_t(j) = \max_{1 \leq i \leq N} \{a_{ij} \delta_{t-1}(i)\} b_j(x_t)$, $2 \leq t \leq T$ and $1 \leq j \leq N$
- $\Psi_t(j) = \text{argmax}_{1 \leq i \leq N} \{a_{ij} \delta_{t-1}(i)\}$, $2 \leq t \leq T$ and $1 \leq j \leq N$

Termination

- $p(X \mid \lambda) = \max_{1 \leq i \leq N} [\delta_T(i)]$

Procedure

Initialization

- $\delta_1(i) = \pi_i b_i(x_1),\ 1 \leq i \leq N$
- Array $\Psi_1(i) = 0\ ,\ 1 \leq i \leq N$

Recusrions

- $\delta_t(j) = \max_{1 \leq i \leq N} \left\{ a_{ij} \delta_{t-1}(i) \right\} b_j(x_t),\ 2 \leq t \leq T$ and $1 \leq j \leq N$
- $\Psi_t(j) = \operatorname{argmax}_{1 \leq i \leq N} \left\{ a_{ij} \delta_{t-1}(i) \right\},\ 2 \leq t \leq T$ and $1 \leq j \leq N$

Termination

- $P* = \max_{1 \leq i \leq N} \left[\delta_T(i) \right]$
- $q_T* = \operatorname{argmax}_{1 \leq i \leq N} \left\{ \delta_T(i) \right\}$

Procedure

Initialization

- $\delta_1(i) = \pi_i b_i(x_1),\ 1 \leq i \leq N$
- Array $\Psi_1(i) = 0\ ,\ 1 \leq i \leq N$

Recusrions

- $\delta_t(j) = \max_{1 \leq i \leq N} \left\{ a_{ij} \delta_{t-1}(i) \right\} b_j(x_t),\ 2 \leq t \leq T$ and $1 \leq j \leq N$
- $\Psi_t(j) = \text{argmax}_{1 \leq i \leq N} \left\{ a_{ij} \delta_{t-1}(i) \right\},\ 2 \leq t \leq T$ and $1 \leq j \leq N$

Termination

- $P* = \max_{1 \leq i \leq N} \left[\delta_T(i) \right]$
- $q_T* = \text{argmax}_{1 \leq i \leq N} \left\{ \delta_T(i) \right\}$

Procedure

Initialization

- $\delta_1(i) = \pi_i b_i(x_1)$, $1 \leq i \leq N$
- Array $\Psi_1(i) = 0$, $1 \leq i \leq N$

Recusrions

- $\delta_t(j) = \max_{1 \leq i \leq N} \left\{ a_{ij} \delta_{t-1}(i) \right\} b_j(x_t)$, $2 \leq t \leq T$ and $1 \leq j \leq N$
- $\Psi_t(j) = \text{argmax}_{1 \leq i \leq N} \left\{ a_{ij} \delta_{t-1}(i) \right\}$, $2 \leq t \leq T$ and $1 \leq j \leq N$

Termination

- $P* = \max_{1 \leq i \leq N} \left[\delta_T(i) \right]$
- $q_T* = \text{argmax}_{1 \leq i \leq N} \left\{ \delta_T(i) \right\}$

Procedure

Initialization

- $\delta_1(i) = \pi_i b_i(x_1)$, $1 \leq i \leq N$
- Array $\Psi_1(i) = 0$, $1 \leq i \leq N$

Recusrions

- $\delta_t(j) = \max_{1 \leq i \leq N} \left[a_{ij} \delta_{t-1}(i) \right] b_j(x_t)$, $2 \leq t \leq T$ and $1 \leq j \leq N$
- $\Psi_t(j) = \text{argmax}_{1 \leq i \leq N} \left[a_{ij} \delta_{t-1}(i) \right]$, $2 \leq t \leq T$ and $1 \leq j \leq N$

Termination

- $P* = \max_{1 \leq i \leq N} \left[\delta_T(i) \right]$
- $q_T* = \text{argmax}_{1 \leq i \leq N} \left[\delta_T(i) \right]$

Clearly

Something Notable

All those recursion can be changed for iterative bottom-up procedure

That all

I leave to you to figure out!!!

Clearly

Something Notable

All those recursion can be changed for iterative bottom-up procedure

That all

I leave to you to figure out!!!

Backtracking

Path backtracking

$q_t* = \Psi_{t+1}(q_{t+1}*)$, $t = T-1, T-2, \cdots, 1$.

Outline

The most difficult

Problem

There is no analytical solution to third problem!!!

However

Given a finite observation sequence as trainign data, we can choose $\lambda(A, B, \pi)$ such that $P(O|\lambda)$is maximized.

For this

We have the Baum-Welch or Forward-Backward Algorithm.

The most difficult

Problem

There is no analytical solution to third problem!!!

However

Given a finite observation sequence as trainign data, we can choose $\lambda(A, B, \pi)$ such that $P(O|\lambda)$is maximized.

For this

We have the Baum-Welch or Forward-Backward Algorithm.

The most difficult

Problem

There is no analytical solution to third problem!!!

However

Given a finite observation sequence as trainign data, we can choose $\lambda(A, B, \pi)$ such that $P(O|\lambda)$is maximized.

For this

We have the Baum-Welch or Forward-Backward Algorithm.

Outline

Another way to see this is thinking in Maximum Likelihood

Given our observed data sequence $X = \langle x_1, x_2, ..., x_T \rangle$

The likelihood function is obtained from the joint distribution by marginalizing over the Hidden State Variables

$$P(X|\lambda) = \sum_{\Omega} P(X, \Omega|\lambda) \tag{30}$$

Problems

- We cannot simply treat each summation over ω_i independently!!!
- We cannot perform all the summations explicitly because it results in N^T terms.

And Again

A further difficulty for the likelihood function is that represents a summation over the emission models for different settings of the latent variables.

Another way to see this is thinking in Maximum Likelihood

Given our observed data sequence $X = \langle x_1, x_2, ..., x_T \rangle$

The likelihood function is obtained from the joint distribution by marginalizing over the Hidden State Variables

$$P(X|\lambda) = \sum_{\Omega} P(X, \Omega|\lambda) \tag{30}$$

Problems

- We cannot simply treat each summation over ω_i independently!!!
- We cannot perform all the summations explicitly because it results in N^T terms.

And Again

A further difficulty for the likelihood function is that represents a summation over the emission models for different settings of the latent variables.

Another way to see this is thinking in Maximum Likelihood

Given our observed data sequence $X = \langle x_1, x_2, ..., x_T \rangle$

The likelihood function is obtained from the joint distribution by marginalizing over the Hidden State Variables

$$P(X|\lambda) = \sum_{\Omega} P(X, \Omega|\lambda) \tag{30}$$

Problems

- We cannot simply treat each summation over ω_i independently!!!
- We cannot perform all the summations explicitly because it results in N^T terms.

And Again

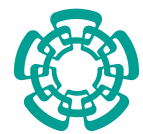
A further difficulty for the likelihood function is that represents a summation over the emission models for different settings of the latent variables.

Outline

Initial Setup

Let us consider discrete (categorical) HMMs of length T (each observation sequence is T observations long).

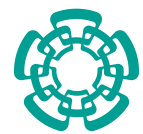
1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. D observations, $\mathcal{X} = \left\{X^{(1)}, X^{(2)}, ..., X^{(D)}\right\}$ such that $X^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, ..., x_T^{(i)}\right\}$.
4. We assume each observation is drawn iid.
5. Finally, we are given $\Omega = \left\{\Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}\right\}$ one sequence of hidden states per observation, such that $\Omega^{(i)} = \left\{\omega_1^{(i)}, \omega_2^{(i)}, ..., \omega_T^{(i)}\right\}$.



Initial Setup

Let us consider discrete (categorical) HMMs of length T (each observation sequence is T observations long).

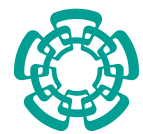
1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. D observations, $\mathcal{X} = \left\{X^{(1)}, X^{(2)}, ..., X^{(D)}\right\}$ such that $X^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, ..., x_T^{(i)}\right\}$.
4. We assume each observation is drawn iid.
5. Finally, we are given $\Omega = \left\{\Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}\right\}$ one sequence of hidden states per observation, such that $\Omega^{(i)} = \left\{\omega_1^{(i)}, \omega_2^{(i)}, ..., \omega_T^{(i)}\right\}$.



Initial Setup

Let us consider discrete (categorical) HMMs of length T (each observation sequence is T observations long).

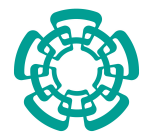
1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. D observations, $\mathcal{X} = \left\{X^{(1)}, X^{(2)}, ..., X^{(D)}\right\}$ such that $X^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, ..., x_T^{(i)}\right\}$.
4. We assume each observation is drawn iid.
5. Finally, we are given $\Omega = \left\{\Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}\right\}$ one sequence of hidden states per observation, such that $\Omega^{(i)} = \left\{\omega_1^{(i)}, \omega_2^{(i)}, ..., \omega_T^{(i)}\right\}$.



Initial Setup

Let us consider discrete (categorical) HMMs of length T (each observation sequence is T observations long).

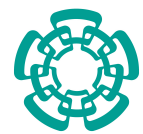
1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. D observations, $\mathcal{X} = \left\{X^{(1)}, X^{(2)}, ..., X^{(D)}\right\}$ such that $X^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, ..., x_T^{(i)}\right\}$.
4. We assume each observation is drawn iid.
5. Finally, we are given $\Omega = \left\{\Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}\right\}$ one sequence of hidden states per observation, such that $\Omega^{(i)} = \left\{\omega_1^{(i)}, \omega_2^{(i)}, ..., \omega_T^{(i)}\right\}$.



Initial Setup

Let us consider discrete (categorical) HMMs of length T (each observation sequence is T observations long).

1. N is the number of hidden states.
2. M is the number of different observation symbols.
3. D observations, $\mathcal{X} = \left\{X^{(1)}, X^{(2)}, ..., X^{(D)}\right\}$ such that $X^{(i)} = \left\{x_1^{(i)}, x_2^{(i)}, ..., x_T^{(i)}\right\}$.
4. We assume each observation is drawn iid.
5. Finally, we are given $\mathbf{\Omega} = \left\{\Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}\right\}$ one sequence of hidden states per observation, such that $\Omega^{(i)} = \left\{\omega_1^{(i)}, \omega_2^{(i)}, ..., \omega_T^{(i)}\right\}$.



Outline

Baum-Welch Algorithm

Remember the EM Algorithm

1. Compute $Q(\lambda, \lambda^n) = \sum_{\mathbf{\Omega}} \log [P(\mathcal{X}, \mathbf{\Omega}|\lambda)] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n)$
2. Set $\lambda^{n+1} = \arg\max_{\lambda} Q(\lambda, \lambda^n)$

Baum-Welch Algorithm

Remember the EM Algorithm

1. Compute $Q(\lambda, \lambda^n) = \sum_{\mathbf{\Omega}} \log\left[P(\mathcal{X}, \mathbf{\Omega}|\lambda)\right] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n)$
2. Set $\lambda^{n+1} = \arg\max_{\lambda} Q(\lambda, \lambda^n)$

We can rewrite the previous iterations (Given $P(X, \Omega) = P(X)P(\Omega|X)$) as

$$\arg\max_{\lambda} \sum_{\Omega} \log\left[P(\mathcal{X}, \Omega|\lambda)\right] P(\Omega|\mathcal{X}, \lambda^n) = \arg\max_{\lambda} \sum_{\Omega} \log\left[P(\mathcal{X}, \Omega|\lambda)\right] \frac{P(\Omega, \mathcal{X}|\lambda^n)}{P(\mathcal{X}, |\lambda^n)}$$

$$\approx \arg\max_{\lambda} \sum_{\Omega} \log\left[P(\mathcal{X}, \Omega|\lambda)\right] P(\Omega, \mathcal{X}|\lambda^n)$$

$$= \arg\max_{\lambda} \hat{Q}(\lambda, \lambda^n)$$

Baum-Welch Algorithm

Remember the EM Algorithm

1. Compute $Q(\lambda, \lambda^n) = \sum_{\mathbf{\Omega}} \log[P(\mathcal{X}, \mathbf{\Omega}|\lambda)] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n)$
2. Set $\lambda^{n+1} = \arg\max_{\lambda} Q(\lambda, \lambda^n)$

We can rewrite the previous iterations (Given $P(X, \Omega) = P(X)P(\Omega|X)$) as

$$\arg\max_{\lambda} \sum_{\Omega} \log[P(\mathcal{X}, \mathbf{\Omega}|\lambda)] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n) = \arg\max_{\lambda} \sum_{\Omega} \log[P(\mathcal{X}, \mathbf{\Omega}|\lambda)] \frac{P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)}{P(\mathcal{X}, |\lambda^n)}$$

$$\approx \arg\max_{\lambda} \sum_{\Omega} \log[P(\mathcal{X}, \Omega|\lambda)] P(\Omega, \mathcal{X}|\lambda^n)$$

$$= \arg\max_{\lambda} \hat{Q}(\lambda, \lambda^n)$$

Baum-Welch Algorithm

Remember the EM Algorithm

1. Compute $Q(\lambda, \lambda^n) = \sum_{\mathbf{\Omega}} \log\left[P(\mathcal{X}, \mathbf{\Omega}|\lambda)\right] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n)$
2. Set $\lambda^{n+1} = \arg\max_{\lambda} Q(\lambda, \lambda^n)$

We can rewrite the previous iterations (Given $P(X, \Omega) = P(X)P(\Omega|X)$) as

$$\arg\max_{\lambda} \sum_{\Omega} \log\left[P(\mathcal{X}, \mathbf{\Omega}|\lambda)\right] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n) = \arg\max_{\lambda} \sum_{\Omega} \log\left[P(\mathcal{X}, \mathbf{\Omega}|\lambda)\right] \frac{P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)}{P(\mathcal{X}, |\lambda^n)}$$

$$\approx \arg\max_{\lambda} \sum_{\Omega} \log\left[P(X, \mathbf{\Omega}|\lambda)\right] P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)$$

$$= \arg\max_{\lambda} \hat{Q}(\lambda, \lambda^n)$$

Baum-Welch Algorithm

Remember the EM Algorithm

1. Compute $Q(\lambda, \lambda^n) = \sum_{\mathbf{\Omega}} \log[P(\mathcal{X}, \mathbf{\Omega}|\lambda)] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n)$
2. Set $\lambda^{n+1} = \arg\max_{\lambda} Q(\lambda, \lambda^n)$

We can rewrite the previous iterations (Given $P(X, \Omega) = P(X)P(\Omega|X)$) as

$$\begin{aligned}
\arg\max_{\lambda} \sum_{\Omega} \log[P(\mathcal{X}, \mathbf{\Omega}|\lambda)] P(\mathbf{\Omega}|\mathcal{X}, \lambda^n) &= \arg\max_{\lambda} \sum_{\Omega} \log[P(\mathcal{X}, \mathbf{\Omega}|\lambda)] \frac{P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)}{P(\mathcal{X}, |\lambda^n)} \\
&\approx \arg\max_{\lambda} \sum_{\Omega} \log[P(X, \mathbf{\Omega}|\lambda)] P(\mathbf{\Omega}, \mathcal{X}|\lambda^n) \\
&= \arg\max_{\lambda} \hat{Q}(\lambda, \lambda^n)
\end{aligned}$$

Cinvestav

After all

We know that

$P(\mathcal{X}|\lambda^n)$ is not effected by λ^n!!!

Now, assuming we have D observations

After all

We know that

$P(\mathcal{X}|\lambda^n)$ is not effected by λ^n!!!

Now, assuming we have D observations

$$P(\mathcal{X}, \mathbf{\Omega}|\lambda) = P\left(X^{(1)}, X^{(2)}, ..., X^{(D)}, \Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}|\lambda\right)$$

$$= P\left(X^{(1)}, \Omega^{(1)}, X^{(2)}, \Omega^{(2)}, ..., X^{(D)}, \Omega^{(D)}|\lambda\right)$$

$$= \prod_{d=1}^{D} P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}, \omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right)$$

$$= \prod_{d=1}^{D} \left[P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}|\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}, \lambda\right) \times ...\right.$$

$$\left. P\left(\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right)\right]$$

After all

We know that

$P(\mathcal{X}|\lambda^n)$ is not effected by λ^n!!!

Now, assuming we have D observations

$$\begin{aligned} P(\mathcal{X}, \mathbf{\Omega}|\lambda) =& P\left(X^{(1)}, X^{(2)}, ..., X^{(D)}, \Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}|\lambda\right) \\ =& P\left(X^{(1)}, \Omega^{(1)}, X^{(2)}, \Omega^{(2)}, ..., X^{(D)}, \Omega^{(D)}|\lambda\right) \end{aligned}$$

$$= \prod_{d=1}^{D} P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}, \omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right)$$

$$= \prod_{d=1}^{D} \left[P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}|\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}, \lambda\right) \times ... \right.$$

$$\left. P\left(\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right)\right]$$

After all

We know that

$P(\mathcal{X}|\lambda^n)$ is not effected by λ^n!!!

Now, assuming we have D observations

$$\begin{aligned}
P(\mathcal{X}, \mathbf{\Omega}|\lambda) =& P\left(X^{(1)}, X^{(2)}, ..., X^{(D)}, \Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}|\lambda\right) \\
=& P\left(X^{(1)}, \Omega^{(1)}, X^{(2)}, \Omega^{(2)}, ..., X^{(D)}, \Omega^{(D)}|\lambda\right) \\
=& \prod_{d=1}^{D} P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}, \omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right)
\end{aligned}$$

$$= \prod_{d=1}^{D} \left[P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}|\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}, \lambda\right) \times ... \right.$$

$$\left. P\left(\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right) \right]$$

After all

We know that

$P(\mathcal{X}|\lambda^n)$ is not effected by λ^n!!!

Now, assuming we have D observations

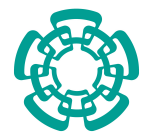
$$
\begin{aligned}
P(\mathcal{X}, \mathbf{\Omega}|\lambda) =& P\left(X^{(1)}, X^{(2)}, ..., X^{(D)}, \Omega^{(1)}, \Omega^{(2)}, ..., \Omega^{(D)}|\lambda\right) \\
=& P\left(X^{(1)}, \Omega^{(1)}, X^{(2)}, \Omega^{(2)}, ..., X^{(D)}, \Omega^{(D)}|\lambda\right) \\
=& \prod_{d=1}^{D} P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}, \omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right) \\
=& \prod_{d=1}^{D} \left[P\left(x_1^{(d)}, x_2^{(d)}, ..., x_T^{(d)}|\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}, \lambda\right) \times ... \right. \\
& \left. P\left(\omega_1^{(d)}, \omega_2^{(d)}, ..., \omega_T^{(d)}|\lambda\right)\right]
\end{aligned}
$$

Thus

We have

$$P\left(\mathcal{X}, \mathbf{\Omega}|\lambda\right) = \prod_{d=1}^{D}\left[\prod_{t=1}^{T} P\left(x_t^{(d)}|\omega_t^{(d)}, \lambda\right)\left[\prod_{t=2}^{T} P\left(\omega_t^{(d)}|\omega_{t-1}^{(d)}, \lambda\right)\right] P\left(\omega_1^{(d)}|\lambda\right)\right]$$

$$= \prod_{d=1}^{D}\left[\pi_{q_1^{(d)}} b_{q_1^{(d)}}\left(x_1^{(d)}\right)\left[\prod_{t=2}^{T} a_{q_{t-1}^{(d)} q_t^{(d)}} b_{q_t^{(d)}}\left(x_t^{(d)}\right)\right]\right]$$

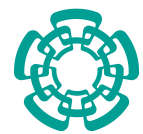


Thus

We have

$$P\left(\mathcal{X}, \mathbf{\Omega}|\lambda\right) = \prod_{d=1}^{D}\left[\prod_{t=1}^{T} P\left(x_t^{(d)}|\omega_t^{(d)}, \lambda\right)\left[\prod_{t=2}^{T} P\left(\omega_t^{(d)}|\omega_{t-1}^{(d)}, \lambda\right)\right] P\left(\omega_1^{(d)}|\lambda\right)\right]$$

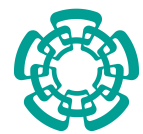
$$= \prod_{d=1}^{D}\left[\pi_{q_1^{(d)}} b_{q_1^{(d)}}\left(x_1^{(d)}\right)\left[\prod_{t=2}^{T} a_{q_{t-1}^{(d)} q_t^{(d)}} b_{q_t^{(d)}}\left(x_t^{(d)}\right)\right]\right]$$



Now, taking the logarithm

We get

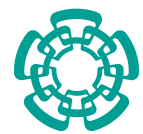
$$\log P(\mathcal{X}, \mathbf{\Omega}|\lambda) = \sum_{d=1}^{D} \left[\log \pi_{q_1^{(d)}} + \ldots \right.$$
$$\sum_{t=2}^{T} \log a_{q_{t-1}^{(d)} q_t^{(d)}} + \ldots$$
$$\left. \sum_{t=1}^{T} \log b_{q_t^{(d)}} \left(x_t^{(d)} \right) \right]$$



We plug back into $\hat{Q}(\lambda, \lambda^n)$

We get the following

$$\begin{aligned}\hat{Q}(\lambda, \lambda^n) = &\sum_{\mathbf{\Omega}} \sum_{d=1}^{D} \log \pi_{q_1^{(d)}} P(\mathbf{\Omega}, \mathcal{X}|\lambda^n) \\ &\sum_{\mathbf{\Omega}} \sum_{d=1}^{D} \sum_{t=2}^{T} \log a_{q_{t-1}^{(d)} q_t^{(d)}} P(\mathbf{\Omega}, \mathcal{X}|\lambda^n) + ... \\ &\sum_{\mathbf{\Omega}} \sum_{d=1}^{D} \sum_{t=1}^{T} \log b_{q_t^{(d)}}\left(x_t^{(d)}\right) P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\end{aligned}$$



Ok

We have reintroduced $P(\Omega, \mathcal{X}|\lambda^n)$

- It looks like it will increase the complexity of our calculations

However

Using marginalization, we will be able to remove the terms that are not necessary to calculate the necessary updates.

Which Updates?

π_i , a_{ij} and $b_j(k)$

Ok

We have reintroduced $P(\Omega, \mathcal{X}|\lambda^n)$

- It looks like it will increase the complexity of our calculations

However

Using marginalization, we will be able to remove the terms that are not necessary to calculate the necessary updates.

Which Updates?

π_i , a_{ij} and $b_j(k)$

Ok

We have reintroduced $P(\Omega, \mathcal{X}|\lambda^n)$

- It looks like it will increase the complexity of our calculations

However

Using marginalization, we will be able to remove the terms that are not necessary to calculate the necessary updates.

Which Updates?

π_i , a_{ij} and $b_j(k)$

Outline

Now, Which Lagrange Multipliers?

One for the M hidden marginal states of π

$$\sum_{i=1}^{N} \pi_i = 1 \tag{31}$$

One for each i

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{32}$$

One for each i

$$\sum_{k=1}^{M} b_i(k) = 1 \tag{33}$$

Remember: $b_i(k) = P(x_k | q_t = \omega_i)$

Now, Which Lagrange Multipliers?

One for the M hidden marginal states of π

$$\sum_{i=1}^{N} \pi_i = 1 \tag{31}$$

One for each i

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{32}$$

One for each i

$$\sum_{k=1}^{M} b_i(k) = 1 \tag{33}$$

Remember: $b_i(k) = P(x_k | q_t = \omega_i)$

Now, Which Lagrange Multipliers?

One for the M hidden marginal states of π

$$\sum_{i=1}^{N} \pi_i = 1 \tag{31}$$

One for each i

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{32}$$

One for each i

$$\sum_{k=1}^{M} b_i(k) = 1 \tag{33}$$

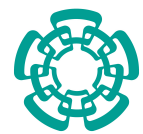
Remember: $b_i(k) = P(x_k | q_t = \omega_i)$

Finally, we have the new Lagrangian

We have

$$\hat{L}\left(\lambda,\lambda^{n}\right)=\hat{Q}\left(\lambda,\lambda^{n}\right)-\lambda_{\pi}\left(\sum_{i=1}^{N}\pi_{i}-1\right)-\sum_{i=1}^{N}\lambda_{a_{i}}\left(\sum_{j=1}^{N}a_{ij}-1\right)-\sum_{i=1}^{N}\lambda_{b_{i}}\left(\sum_{k=1}^{M}b_{i}\left(k\right)-1\right)$$

Outline



Now, if we derive with respect to π_i

We have that

$$\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i}\left(\sum_{\mathbf{\Omega}}\sum_{d=1}^{D} \log \pi_{q_1^{(d)}} P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\right) - \lambda_\pi$$

$$= \frac{\partial}{\partial \pi_i}\left(\sum_{j=1}^{N}\sum_{d=1}^{D} \log \pi_j P\left(q_1^{(d)} = j, \mathcal{X}|\lambda^n\right)\right) - \lambda_\pi$$

$$= \frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X}|\lambda^n\right) - \lambda_\pi = 0$$

Remark: The second step is simply the result of marginalizing out, for each d, all $q_{t\neq 1}^{(d)}$ and $q_t^{(d'\neq d)}$ for all t.

Now, if we derive with respect to π_i

We have that

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial \pi_i} &= \frac{\partial}{\partial \pi_i}\left(\sum_{\mathbf{\Omega}}\sum_{d=1}^{D} \log \pi_{q_1^{(d)}} P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\right) - \lambda_\pi \\ &= \frac{\partial}{\partial \pi_i}\left(\sum_{j=1}^{N}\sum_{d=1}^{D} \log \pi_j P\left(q_1^{(d)} = j, \mathcal{X}|\lambda^n\right)\right) - \lambda_\pi\end{aligned}$$

$$= \frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X}|\lambda^n\right) - \lambda_\pi = 0$$

Remark: The second step is simply the result of marginalizing out, for each d, all $q_{t\neq 1}^{(d)}$ and $q_t^{(d'\neq d)}$ for all t.

Now, if we derive with respect to π_i

We have that

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial \pi_i} &= \frac{\partial}{\partial \pi_i}\left(\sum_{\mathbf{\Omega}}\sum_{d=1}^{D} \log \pi_{q_1^{(d)}} P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\right) - \lambda_\pi \\ &= \frac{\partial}{\partial \pi_i}\left(\sum_{j=1}^{N}\sum_{d=1}^{D} \log \pi_j P\left(q_1^{(d)} = j, \mathcal{X}|\lambda^n\right)\right) - \lambda_\pi \\ &= \frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X}|\lambda^n\right) - \lambda_\pi = 0\end{aligned}$$

Remark: The second step is simply the result of marginalizing out, for each d, all $q_{t\neq 1}^{(d)}$ and $q_t^{(d'\neq d)}$ for all t.

Now, if we derive with respect to π_i

We have that

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial \pi_i} &= \frac{\partial}{\partial \pi_i}\left(\sum_{\mathbf{\Omega}}\sum_{d=1}^{D} \log \pi_{q_1^{(d)}} P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\right) - \lambda_\pi \\ &= \frac{\partial}{\partial \pi_i}\left(\sum_{j=1}^{N}\sum_{d=1}^{D} \log \pi_j P\left(q_1^{(d)} = j, \mathcal{X}|\lambda^n\right)\right) - \lambda_\pi \\ &= \frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X}|\lambda^n\right) - \lambda_\pi = 0\end{aligned}$$

Remark: The second step is simply the result of marginalizing out, for each d, all $q_{t\neq 1}^{(d)}$ and $q_t^{(d'\neq d)}$ for all t.

Now, if we derive with respect to π_i

After all

$$\sum_{d=1}^{D} \log \pi_{q_1^{(d)}} \sum_{\Omega} P\left(\mathbf{\Omega}, \mathcal{X}|\lambda^n\right) = \sum_{d=1}^{D} \log \pi_{q_1^{(d)}} \sum_{j_1^1=1}^{N} \sum_{j_2^1=1}^{N} \cdots \sum_{j_T^1=1}^{N} \cdots \sum_{j_1^T=1}^{N} \sum_{j_2^T=1}^{N} \cdots \sum_{j_T^T=1}^{N} P\left(\mathbf{\Omega}, \mathcal{X}|\lambda^n\right)$$

Now, if we derive with respect to π_i

After all

$$\sum_{d=1}^{D} \log \pi_{q_1^{(d)}} \sum_{\Omega} P\left(\mathbf{\Omega}, \mathcal{X}|\lambda^n\right) = \sum_{d=1}^{D} \log \pi_{q_1^{(d)}} \sum_{j_1^1=1}^{N}\sum_{j_2^1=1}^{N} \dots \sum_{j_T^1=1}^{N} \dots \sum_{j_1^T=1}^{N}\sum_{j_2^T=1}^{N} \dots \sum_{j_T^T=1}^{N} P\left(\mathbf{\Omega}, \mathcal{X}|\lambda^n\right)$$

$$= \sum_{d=1}^{D} \log \pi_{q_1^{(d)}} \sum_{j_1^1=1}^{N}\sum_{j_1^2=1}^{N} \dots \sum_{j_1^T=1}^{N} P\left(q_1^{(1)} = j_1^1, \dots, q_T^{(T)} = j_1^T, \mathcal{X}|\lambda^n\right)$$

$$= \sum_{d=1}^{D} \log \pi_{q_1^{(d)}} \sum_{j_1=1}^{N} \left(q_1^{(d)} = j_1, \mathcal{X}|\lambda^n\right)$$

Thus we have

$$\pi_i = \frac{1}{\lambda_\pi} \sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X}|\lambda^n\right) \tag{34}$$

Now, if we derive with respect to π_i

After all

$$\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{\Omega}P\left(\mathbf{\Omega},\mathcal{X}|\lambda^n\right)=\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{j_1^1=1}^{N}\sum_{j_2^1=1}^{N}\cdots\sum_{j_T^1=1}^{N}\cdots\sum_{j_1^T=1}^{N}\sum_{j_2^T=1}^{N}\cdots\sum_{j_T^T=1}^{N}P\left(\mathbf{\Omega},\mathcal{X}|\lambda^n\right)$$

$$=\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{j_1^1=1}^{N}\sum_{j_1^2=1}^{N}\cdots\sum_{j_1^T=1}^{N}P\left(q_1^{(1)}=j_1^1,\ldots,q_T^{(T)}=j_1^T,\mathcal{X}|\lambda^n\right)$$

$$=\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{j_1=1}^{N}\left(q_1^{(d)}=j_1,\mathcal{X}|\lambda^n\right)$$

Thus we have

$$\pi_i=\frac{1}{\lambda_\pi}\sum_{d=1}^{D}P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)\tag{34}$$

Now, if we derive with respect to π_i

After all

$$\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{\Omega}P\left(\mathbf{\Omega},\mathcal{X}|\lambda^n\right)=\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{j_1^1=1}^{N}\sum_{j_2^1=1}^{N}\cdots\sum_{j_T^1=1}^{N}\cdots\sum_{j_1^T=1}^{N}\sum_{j_2^T=1}^{N}\cdots\sum_{j_T^T=1}^{N}P\left(\mathbf{\Omega},\mathcal{X}|\lambda^n\right)$$

$$=\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{j_1^1=1}^{N}\sum_{j_1^2=1}^{N}\cdots\sum_{j_1^T=1}^{N}P\left(q_1^{(1)}=j_1^1,\ldots,q_T^{(T)}=j_1^T,\mathcal{X}|\lambda^n\right)$$

$$=\sum_{d=1}^{D}\log\pi_{q_1^{(d)}}\sum_{j_1=1}^{N}\left(q_1^{(d)}=j_1,\mathcal{X}|\lambda^n\right)$$

Thus we have

$$\pi_i=\frac{1}{\lambda_\pi}\sum_{d=1}^{D}P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)\tag{34}$$

Now deriving against λ_π

We have that

$$\sum_{i=1}^{N} \pi_i = 1 \tag{35}$$

Thus, substituting into the past equation

$$\frac{1}{\lambda_\pi} \sum_{i=1}^{N} \sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X} | \lambda^n\right) = 1 \tag{36}$$

Then

$$\lambda_\pi = \sum_{i=1}^{N} \sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X} | \lambda^n\right) \tag{37}$$

Now deriving against λ_π

We have that

$$\sum_{i=1}^{N} \pi_i = 1 \tag{35}$$

Thus, substituting into the past equation

$$\frac{1}{\lambda_\pi} \sum_{i=1}^{N} \sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X} | \lambda^n\right) = 1 \tag{36}$$

Then

$$\lambda_\pi = \sum_{i=1}^{N} \sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X} | \lambda^n\right) \tag{37}$$

Now deriving against λ_π

We have that

$$\sum_{i=1}^{N} \pi_i = 1 \tag{35}$$

Thus, substituting into the past equation

$$\frac{1}{\lambda_\pi} \sum_{i=1}^{N} \sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X} | \lambda^n\right) = 1 \tag{36}$$

Then

$$\lambda_\pi = \sum_{i=1}^{N} \sum_{d=1}^{D} P\left(q_1^{(d)} = i, \mathcal{X} | \lambda^n\right) \tag{37}$$

Cinvestav

Plugging back

We get

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)-\lambda_\pi=0$$

Plugging back

We get

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)-\lambda_\pi=0$$

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)=\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right) P\left(q_1^{(d)}=i|\mathcal{X},\lambda^n\right)}{DP\left(\mathcal{X}|\lambda^n\right)}$$

Plugging back

We get

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)-\lambda_\pi=0$$

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)=\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right) P\left(q_1^{(d)}=i|\mathcal{X},\lambda^n\right)}{DP\left(\mathcal{X}|\lambda^n\right)}$$

Plugging back

We get

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)-\lambda_\pi=0$$

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)=\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right) P\left(q_1^{(d)}=i|\mathcal{X},\lambda^n\right)}{DP\left(\mathcal{X}|\lambda^n\right)}$$

Plugging back

We get

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)-\lambda_\pi=0$$

$$\frac{1}{\pi_i}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)=\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{i=1}^{N}\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(q_1^{(d)}=i,\mathcal{X}|\lambda^n\right)}{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right)}$$

$$\pi_i=\frac{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right) P\left(q_1^{(d)}=i|\mathcal{X},\lambda^n\right)}{DP\left(\mathcal{X}|\lambda^n\right)}$$

In addition

We can modify our γ

$$\gamma_{q_{t-1}^{(d)}}(i) = P\left(q_{t-1}^{(d)} = i | X^{(d)}, \lambda^n\right) \tag{38}$$

Meaning

The probability of being in state i at time $q_{t-1}^{(d)}$ given the observation sequence $X^{(d)}$ and the model λ^n.

In addition

We can modify our γ

$$\gamma_{q_{t-1}^{(d)}}(i) = P\left(q_{t-1}^{(d)} = i | X^{(d)}, \lambda^n\right) \tag{38}$$

Meaning

The probability of being in state i at time $q_{t-1}^{(d)}$ given the observation sequence $X^{(d)}$ and the model λ^n.

Thus

We have

$$\pi_i = \frac{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right) P\left(q_1^{(d)} = i|\mathcal{X}, \lambda^n\right)}{DP\left(\mathcal{X}|\lambda^n\right)}$$

$$\pi_i = \frac{1}{D}\sum_{d=1}^{D} P\left(q_1^{(d)} = i|X^{(d)}, \lambda^n\right)$$

$$\pi_i = \frac{1}{D}\sum_{d=1}^{D} \gamma_{q_1^{(d)}}(i)$$

Remark: This can be thought as the expected frequency (Number of Times) in state i at time $(t = 1) = \gamma_{q_1^{(d)}}(i)$

Thus

We have

$$\pi_i = \frac{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right) P\left(q_1^{(d)} = i|\mathcal{X}, \lambda^n\right)}{DP\left(\mathcal{X}|\lambda^n\right)}$$

$$\pi_i = \frac{1}{D}\sum_{d=1}^{D} P\left(q_1^{(d)} = i|X^{(d)}, \lambda^n\right)$$

$$\pi_i = \frac{1}{D}\sum_{d=1}^{D} \gamma_{q_1^{(d)}}(i)$$

Remark: This can be thought as the expected frequency (Number of Times) in state i at time $(t = 1) = \gamma_{q_1^{(d)}}(i)$

Thus

We have

$$\pi_i = \frac{\sum_{d=1}^{D} P\left(\mathcal{X}|\lambda^n\right) P\left(q_1^{(d)} = i|\mathcal{X}, \lambda^n\right)}{DP\left(\mathcal{X}|\lambda^n\right)}$$

$$\pi_i = \frac{1}{D}\sum_{d=1}^{D} P\left(q_1^{(d)} = i|X^{(d)}, \lambda^n\right)$$

$$\pi_i = \frac{1}{D}\sum_{d=1}^{D} \gamma_{q_1^{(d)}}\left(i\right)$$

Remark: This can be thought as the expected frequency (Number of Times) in state i at time $(t = 1) = \gamma_{q_1^{(d)}}\left(i\right)$

Now, for a_{ij}

We now follow a similar process for the a_{ij}

$$\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial a_{ij}} = \frac{\partial}{\partial a_{ij}} \left(\sum_{\mathbf{\Omega}} \sum_{d=1}^{D} \sum_{t=2}^{T} \log a_{q_{t-1}^{(d)} q_t^{(d)}} P(\mathbf{\Omega}, \mathcal{X} | \lambda^n) \right) - \sum_{i=1}^{N} \lambda_{a_i}$$

$$= \frac{\partial}{\partial a_{ij}} \left(\sum_{h=1}^{N} \sum_{k=1}^{N} \sum_{d=1}^{D} \sum_{t=2}^{T} \log a_{jk} P\left(q_{t-1}^{(d)} = h, q_t^{(d)} = k, \mathcal{X} | \lambda^n\right) \right) - \lambda_{a_i}$$

$$= \frac{1}{a_{ij}} \sum_{d=1}^{D} \sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X} | \lambda^n\right) - \lambda_{a_i} = 0$$

Now, for a_{ij}

We now follow a similar process for the a_{ij}

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial a_{ij}} &= \frac{\partial}{\partial a_{ij}}\left(\sum_{\mathbf{\Omega}}\sum_{d=1}^{D}\sum_{t=2}^{T}\log a_{q_{t-1}^{(d)}q_t^{(d)}} P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\right) - \sum_{i=1}^{N}\lambda_{a_i} \\ &= \frac{\partial}{\partial a_{ij}}\left(\sum_{h=1}^{N}\sum_{k=1}^{N}\sum_{d=1}^{D}\sum_{t=2}^{T}\log a_{jk} P\left(q_{t-1}^{(d)} = h, q_t^{(d)} = k, \mathcal{X}|\lambda^n\right)\right) - \lambda_{a_i}\end{aligned}$$

$$= \frac{1}{a_{ij}}\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right) - \lambda_{a_i} = 0$$

In a similar way

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{39}$$

Now, for a_{ij}

We now follow a similar process for the a_{ij}

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial a_{ij}} &= \frac{\partial}{\partial a_{ij}}\left(\sum_{\mathbf{\Omega}}\sum_{d=1}^{D}\sum_{t=2}^{T}\log a_{q_{t-1}^{(d)}q_t^{(d)}}P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\right) - \sum_{i=1}^{N}\lambda_{a_i} \\ &= \frac{\partial}{\partial a_{ij}}\left(\sum_{h=1}^{N}\sum_{k=1}^{N}\sum_{d=1}^{D}\sum_{t=2}^{T}\log a_{jk}P\left(q_{t-1}^{(d)} = h, q_t^{(d)} = k, \mathcal{X}|\lambda^n\right)\right) - \lambda_{a_i} \\ &= \frac{1}{a_{ij}}\sum_{d=1}^{D}\sum_{t=2}^{T}P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right) - \lambda_{a_i} = 0\end{aligned}$$

In a similar way

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{39}$$

Now, for a_{ij}

We now follow a similar process for the a_{ij}

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial a_{ij}} &= \frac{\partial}{\partial a_{ij}}\left(\sum_{\mathbf{\Omega}}\sum_{d=1}^{D}\sum_{t=2}^{T}\log a_{q_{t-1}^{(d)}q_t^{(d)}}P(\mathbf{\Omega}, \mathcal{X}|\lambda^n)\right) - \sum_{i=1}^{N}\lambda_{a_i} \\ &= \frac{\partial}{\partial a_{ij}}\left(\sum_{h=1}^{N}\sum_{k=1}^{N}\sum_{d=1}^{D}\sum_{t=2}^{T}\log a_{jk}P\left(q_{t-1}^{(d)} = h, q_t^{(d)} = k, \mathcal{X}|\lambda^n\right)\right) - \lambda_{a_i} \\ &= \frac{1}{a_{ij}}\sum_{d=1}^{D}\sum_{t=2}^{T}P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right) - \lambda_{a_i} = 0\end{aligned}$$

In a similar way

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{39}$$

Then

We have that

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right)}{\lambda_{a_i}}$$

Then

We have that

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right)}{\lambda_{a_i}}$$

Then, using $\sum_{j=1}^{N} a_{ij} = 1$

$$\lambda_{a_i} = \sum_{j=1}^{N}\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right) \tag{40}$$

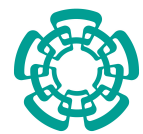
Re-expressing

We get

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_{t}^{(d)} = j, \mathcal{X}|\lambda^n\right)}{\sum_{j=1}^{N}\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_{t}^{(d)} = j, \mathcal{X}|\lambda^n\right)}$$

Marginalizing the denominator and numerator in $\mathcal{X}$

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_{t}^{(d)} = j, \mathcal{X}|\lambda^n\right)}{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, \mathcal{X}|\lambda^n\right)}$$



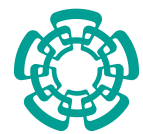
Re-expressing

We get

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right)}{\sum_{j=1}^{N}\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right)}$$

Marginalizing the denominator and numerator in $\mathcal{X}$

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j, \mathcal{X}|\lambda^n\right)}{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, \mathcal{X}|\lambda^n\right)}$$

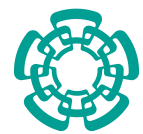


Next

We have

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j|\mathcal{X}, \lambda^n\right) P\left(\mathcal{X}|\lambda^n\right)}{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i|\mathcal{X}, \lambda^n\right) P\left(\mathcal{X}|\lambda^n\right)}$$

$$= \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j|X^{(d)}, \lambda^n\right)}{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i|X^{(d)}, \lambda^n\right)}$$



Now, we are looking for a more compact version

We define

The probability of being in state i at time $q_{t-1}^{(d)}$ and state j at time $q_t^{(d)}$:

$$\xi_{q_{t-1}^{(d)}}(i,j) = P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j | X^{(d)}, \lambda\right) \tag{41}$$

This is Basically a Joint Probability

We have

- A Joint Probability over $q_{t-1}^{(d)} = i$ and $q_t^{(d)} = j$ given data sample $X^{(d)}$

A question raises immediately

- To what is this equivalent?

For this

- We can use Bayes!!!

This is Basically a Joint Probability

We have

- A Joint Probability over $q_{t-1}^{(d)} = i$ and $q_t^{(d)} = j$ given data sample $X^{(d)}$

A question raises immediately

- To what is this equivalent?

For this

- We can use Bayes!!!

This is Basically a Joint Probability

We have

- A Joint Probability over $q_{t-1}^{(d)} = i$ and $q_t^{(d)} = j$ given data sample $X^{(d)}$

A question raises immediately

- To what is this equivalent?

For this

- We can use Bayes!!!

Using Bayes

We have for $\xi_{q_{t-1}^{(d)}}(i,j)$

$$\xi_{q_{t-1}^{(d)}}(i,j) = \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i,q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)}|X^{(d)},\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)},\lambda\right)P\left(q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)}|\lambda\right)}{P\left(X|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(x_{t+1}^{(d)},\ldots,x_T^{(d)}|q_t^{(d)}=j,x_t^{(d)},\lambda\right)\times\cdots}{P\left(X^{(d)},|\lambda\right)}$$

$$\frac{P\left(q_t^{(d)}=j,x_t^{(d)}|\lambda\right)}{}$$

$$= P\left(x_1^{(d)},\cdots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(x_{t+1}^{(d)},\cdots,x_T^{(d)}|q_t^{(d)}=j,x_t^{(d)},\lambda\right)\times\cdots$$

$$\frac{P\left(x_t^{(d)}|q_t^{(d)}=j,\lambda\right)P\left(q_t^{(d)}=j|\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(q_t^{(d)}=j|\lambda\right)\beta_t^{(d)}(j)b_j\left(x_t^{(d)}\right)}{P\left(X^{(d)}|\lambda\right)}=\cdots$$

Using Bayes

We have for $\xi_{q_{t-1}^{(d)}}(i,j)$

$$\xi_{q_{t-1}^{(d)}}(i,j) = \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i,q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)}|X^{(d)},\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)},\lambda\right)P\left(q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)}|\lambda\right)}{P\left(X|\lambda\right)}$$

Using Bayes

We have for $\xi_{q_{t-1}^{(d)}}(i,j)$

$$\xi_{q_{t-1}^{(d)}}(i,j) = \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i,q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)}|X^{(d)},\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)},\lambda\right)P\left(q_t^{(d)}=j,x_t^{(d)},\ldots,x_T^{(d)}|\lambda\right)}{P\left(X|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},\ldots,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(x_{t+1}^{(d)},\ldots,x_T^{(d)}|q_t^{(d)}=j,x_t^{(d)},\lambda\right)\times\cdots}{P\left(X^{(d)},|\lambda\right)}$$

$$\frac{P\left(q_t^{(d)}=j,x_t^{(d)}|\lambda\right)}{}$$

Using Bayes

We have for $\xi_{q_{t-1}^{(d)}}(i,j)$

$$\xi_{q_{t-1}^{(d)}}(i,j) = \frac{P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i,q_t^{(d)}=j,x_t^{(d)},...,x_T^{(d)}|X^{(d)},\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,x_t^{(d)},...,x_T^{(d)},\lambda\right)P\left(q_t^{(d)}=j,x_t^{(d)},...,x_T^{(d)}|\lambda\right)}{P\left(X|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(x_{t+1}^{(d)},...,x_T^{(d)}|q_t^{(d)}=j,x_t^{(d)},\lambda\right)\times\cdots}{P\left(X^{(d)},|\lambda\right)}$$

$$\underline{P\left(q_t^{(d)}=j,x_t^{(d)}|\lambda\right)}$$

$$= P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(x_{t+1}^{(d)},...,x_T^{(d)}|q_t^{(d)}=j,x_t^{(d)},\lambda\right)\times\cdots$$

$$\frac{P\left(x_t^{(d)}|q_t^{(d)}=j,\lambda\right)P\left(q_t^{(d)}=j|\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

Using Bayes

We have for $\xi_{q_{t-1}^{(d)}}(i,j)$

$$\xi_{q_{t-1}^{(d)}}(i,j) = \frac{P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i,q_t^{(d)}=j,x_t^{(d)},...,x_T^{(d)}|X^{(d)},\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,x_t^{(d)},...,x_T^{(d)},\lambda\right)P\left(q_t^{(d)}=j,x_t^{(d)},...,x_T^{(d)}|\lambda\right)}{P\left(X|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(x_{t+1}^{(d)},...,x_T^{(d)}|q_t^{(d)}=j,x_t^{(d)},\lambda\right)\times\cdots}{P\left(X^{(d)},|\lambda\right)}$$

$$\underline{P\left(q_t^{(d)}=j,x_t^{(d)}|\lambda\right)}$$

$$= P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(x_{t+1}^{(d)},...,x_T^{(d)}|q_t^{(d)}=j,x_t^{(d)},\lambda\right)\times\cdots$$

$$\frac{P\left(x_t^{(d)}|q_t^{(d)}=j,\lambda\right)P\left(q_t^{(d)}=j|\lambda\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(x_1^{(d)},...,x_{t-1}^{(d)},q_{t-1}^{(d)}=i|q_t^{(d)}=j,\lambda\right)P\left(q_t^{(d)}=j|\lambda\right)\beta_t^{(d)}(j)\,b_j\left(x_t^{(d)}\right)}{P\left(X^{(d)}|\lambda\right)} = \ldots$$

Again Bayes

Thus $\xi_{q_{t-1}^{(d)}}(i,j) =$

$$= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, q_t^{(d)} = j|\lambda\right) \beta_t(j) b_j(x_t)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(q_t^{(d)} = j|x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, \lambda\right) P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i|\lambda\right) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P(X|\lambda)}$$

$$= \frac{P\left(q_t^{(d)} = j|q_{t-1}^{(d)} = i, \lambda\right) \alpha_{t-1}^{(d)}(i) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{\alpha_{t-1}^{(d)}(i) a_{ij}^{(d)} \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{\sum_{k=1}^{N} \sum_{h=1}^{N} \alpha_{t-1}^{(d)}(k) a_{kh}^{(d)} \beta_t^{(d)}(h) b_h\left(x_t^{(d)}\right)}$$

Again Bayes

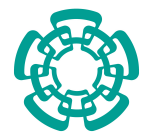
Thus $\xi_{q_{t-1}^{(d)}}(i,j) =$

$$= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, q_t^{(d)} = j|\lambda\right) \beta_t(j) b_j(x_t)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(q_t^{(d)} = j|x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, \lambda\right) P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i|\lambda\right) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P(X|\lambda)}$$

$$= \frac{P\left(q_t^{(d)} = j|q_{t-1}^{(d)} = i, \lambda\right) \alpha_{t-1}^{(d)}(i) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{\alpha_{t-1}^{(d)}(i) a_{ij}^{(d)} \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{\sum_{k=1}^{N}\sum_{h=1}^{N} \alpha_{t-1}^{(d)}(k) a_{kh}^{(d)} \beta_t^{(d)}(h) b_h\left(x_t^{(d)}\right)}$$



Again Bayes

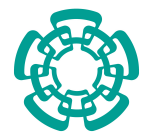
Thus $\xi_{q_{t-1}^{(d)}}(i,j) =$

$$= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, q_t^{(d)} = j|\lambda\right) \beta_t(j) b_j(x_t)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(q_t^{(d)} = j|x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, \lambda\right) P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i|\lambda\right) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P(X|\lambda)}$$

$$= \frac{P\left(q_t^{(d)} = j|q_{t-1}^{(d)} = i, \lambda\right) \alpha_{t-1}^{(d)}(i) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{\alpha_{t-1}^{(d)}(i) a_{ij}^{(d)} \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{\sum_{k=1}^{N}\sum_{h=1}^{N} \alpha_{t-1}^{(d)}(k) a_{kh}^{(d)} \beta_t^{(d)}(h) b_h\left(x_t^{(d)}\right)}$$



Again Bayes

Thus $\xi_{q_{t-1}^{(d)}}(i,j) =$

$$= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, q_t^{(d)} = j|\lambda\right) \beta_t(j) b_j(x_t)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{P\left(q_t^{(d)} = j|x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, \lambda\right) P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i|\lambda\right) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P(X|\lambda)}$$

$$= \frac{P\left(q_t^{(d)} = j|q_{t-1}^{(d)} = i, \lambda\right) \alpha_{t-1}^{(d)}(i) \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{P\left(X^{(d)}|\lambda\right)}$$

$$= \frac{\alpha_{t-1}^{(d)}(i) a_{ij}^{(d)} \beta_t^{(d)}(j) b_j\left(x_t^{(d)}\right)}{\sum_{k=1}^{N}\sum_{h=1}^{N} \alpha_{t-1}^{(d)}(k) a_{kh}^{(d)} \beta_t^{(d)}(h) b_h\left(x_t^{(d)}\right)}$$

Basically an Aggregation

Of the probabilities

- One coming from the Past of sample $X^{(d)}$

$$\text{Information from the Past} = \alpha_{t-1}^{(d)}(i)$$

Ones coming from the Future of sample $X^{(d)}$

$$\text{Information from the Future} = \beta_t^{(d)}(j)$$

Basically an Aggregation

Of the probabilities

- One coming from the Past of sample $X^{(d)}$

$$\text{Information from the Past} = \alpha_{t-1}^{(d)}(i)$$

Ones coming from the Future of sample $X^{(d)}$

$$\text{Information from the Future} = \beta_{t}^{(d)}(j)$$

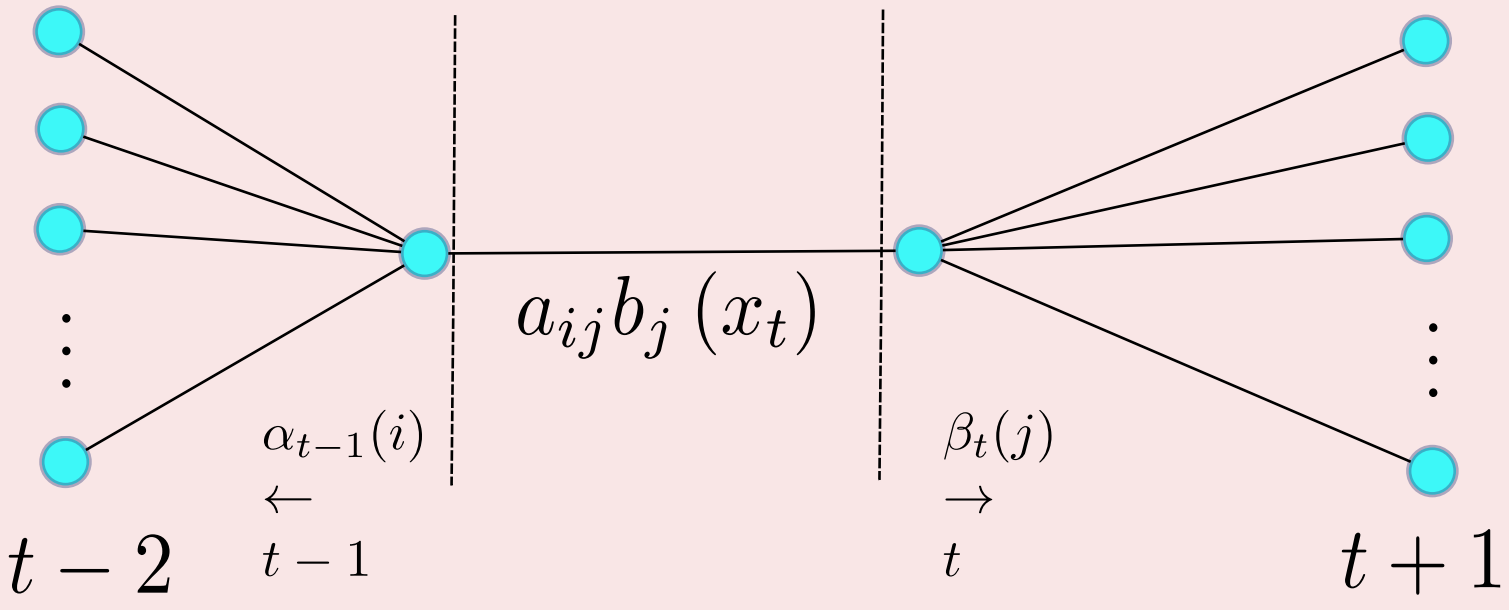
Thus

To reach the symbol $x_t^{(d)}$ at sample $X^{(d)}$

Information when reaching state i and emiting $x_t = a_{ij}^{(d)} b_j \left(x_t^{(d)} \right)$

This can be seen graphically as

This is illustrated by the following figure



In addition

Given the definition of $\gamma_{q_{t-1}^{(d)}}(i)$

$$\gamma_{q_{t-1}^{(d)}}(i) = \frac{\alpha_{q_{t-1}^{(d)}}(i)\,\beta_{q_{t-1}^{(d)}}(i)}{\sum_{j=1}^{N}\alpha_{q_t^{(d)}}(j)\,\beta_{q_t^{(d)}}(j)} \tag{42}$$

Summing over j

In addition

Given the definition of $\gamma_{q_{t-1}^{(d)}}(i)$

$$\gamma_{q_{t-1}^{(d)}}(i) = \frac{\alpha_{q_{t-1}^{(d)}}(i)\,\beta_{q_{t-1}^{(d)}}(i)}{\sum_{j=1}^{N}\alpha_{q_t^{(d)}}(j)\,\beta_{q_t^{(d)}}(j)} \tag{42}$$

Summing over j

$$\sum_{j=1}^{N}\xi_{q_{t-1}^{(d)}}(i,j) = \sum_{j=1}^{N}P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j | X^{(d)}, \lambda\right)$$

$$= P\left(q_{t-1}^{(d)} = i | X^{(d)}, \lambda\right)$$

$$= \frac{P\left(x_1^{(d)}, \cdots, x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, x_t^{(d)}, \cdots, x_T^{(d)} | \lambda\right)}{P\left(X^{(d)} | \lambda\right)}$$

In addition

Given the definition of $\gamma_{q_{t-1}^{(d)}}(i)$

$$\gamma_{q_{t-1}^{(d)}}(i) = \frac{\alpha_{q_{t-1}^{(d)}}(i)\,\beta_{q_{t-1}^{(d)}}(i)}{\sum_{j=1}^{N}\alpha_{q_t^{(d)}}(j)\,\beta_{q_t^{(d)}}(j)} \tag{42}$$

Summing over j

$$\begin{aligned}\sum_{j=1}^{N}\xi_{q_{t-1}^{(d)}}(i,j) &= \sum_{j=1}^{N} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j | X^{(d)}, \lambda\right) \\ &= P\left(q_{t-1}^{(d)} = i | X^{(d)}, \lambda\right)\end{aligned}$$

$$= \frac{P\left(x_1^{(d)}, \cdots, x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, x_t^{(d)}, \cdots, x_T^{(d)} | \lambda\right)}{P\left(X^{(d)} | \lambda\right)}$$

In addition

Given the definition of $\gamma_{q_{t-1}^{(d)}}(i)$

$$\gamma_{q_{t-1}^{(d)}}(i) = \frac{\alpha_{q_{t-1}^{(d)}}(i)\,\beta_{q_{t-1}^{(d)}}(i)}{\sum_{j=1}^{N}\alpha_{q_t^{(d)}}(j)\,\beta_{q_t^{(d)}}(j)} \tag{42}$$

Summing over j

$$\begin{aligned}\sum_{j=1}^{N}\xi_{q_{t-1}^{(d)}}(i,j) &= \sum_{j=1}^{N} P\left(q_{t-1}^{(d)} = i, q_t^{(d)} = j | X^{(d)}, \lambda\right)\\ &= P\left(q_{t-1}^{(d)} = i | X^{(d)}, \lambda\right)\\ &= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, x_t^{(d)}, ..., x_T^{(d)} | \lambda\right)}{P\left(X^{(d)} | \lambda\right)}\end{aligned}$$

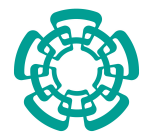
Then

We have

$$= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, x_t^{(d)}, ..., x_T^{(d)} | \lambda\right)}{P\left(X^{(d)} | \lambda\right)}$$

$$= \frac{\alpha_{t-1}(i)\,\beta_{t-1}(i)}{\sum_{i=1}^{N} P\left(x_1, ..., x_t, ..., x_T, q_t = \omega_i | \lambda\right)}$$

$$= \gamma_{q_{t-1}^{(d)}}(i)$$



Then

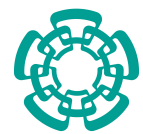
We have

$$
\begin{aligned}
&= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, x_t^{(d)}, ..., x_T^{(d)} | \lambda\right)}{P\left(X^{(d)} | \lambda\right)} \\
&= \frac{\alpha_{t-1}(i)\,\beta_{t-1}(i)}{\sum_{i=1}^{N} P\left(x_1, ..., x_t, ..., x_T, q_t = \omega_i | \lambda\right)} \\
&= \gamma_{q_{t-1}^{(d)}}(i)
\end{aligned}
$$

Then

We have

$$
\begin{aligned}
&= \frac{P\left(x_1^{(d)}, ..., x_{t-1}^{(d)}, q_{t-1}^{(d)} = i, x_t^{(d)}, ..., x_T^{(d)} | \lambda\right)}{P\left(X^{(d)} | \lambda\right)} \\
&= \frac{\alpha_{t-1}(i)\,\beta_{t-1}(i)}{\sum_{i=1}^{N} P\left(x_1, ..., x_t, ..., x_T, q_t = \omega_i | \lambda\right)} \\
&= \gamma_{q_{t-1}^{(d)}}(i)
\end{aligned}
$$



We can also think about the expected values

The following sum

1. It can be thought as expected (over time) number of times that state ω_i is visited
2. Or the number of transitions made from state ω_i

It is

$$\sum_{t=2}^{T} \gamma_{q_{t-1}^{(d)}}(i) \tag{43}$$

We can also think about the expected values

The following sum

1. It can be thought as expected (over time) number of times that state ω_i is visited
2. Or the number of transitions made from state ω_i

It is

$$\sum_{t=2}^{T} \gamma_{q_{t-1}^{(d)}}(i) \tag{43}$$

Also the expected value for $\xi_{q_{t-1}^{(d)}}(i,j)$

The following sum

It can be thought as the expected number of transitions from i to j

It is

$$\sum_{t=2}^{T} \xi_{q_{t-1}^{(d)}}(i,j) \tag{44}$$

Also the expected value for $\xi_{q_{t-1}^{(d)}}(i,j)$

The following sum

It can be thought as the expected number of transitions from i to j

It is

$$\sum_{t=2}^{T} \xi_{q_{t-1}^{(d)}}(i,j) \tag{44}$$

Remark

Important

Baum et al. proved that the maximization of$Q\left(\lambda,\overline{\lambda}\right)$ leads to increased likelihood:

$$\max_{\overline{\lambda}}\left[Q\left(\lambda,\overline{\lambda}\right)\right]\Rightarrow P\left(X|\overline{\lambda}\right)\geq P\left(X|\lambda\right) \tag{45}$$

For details

Please look back at our slides on EM!!!

Remark

Important

Baum et al. proved that the maximization of$Q\left(\lambda,\overline{\lambda}\right)$ leads to increased likelihood:

$$\max_{\overline{\lambda}}\left[Q\left(\lambda,\overline{\lambda}\right)\right]\Rightarrow P\left(X|\overline{\lambda}\right)\geq P\left(X|\lambda\right) \tag{45}$$

For details

Please look back at our slides on EM!!!

Remark

Important

Baum et al. proved that the maximization of$Q\left(\lambda,\overline{\lambda}\right)$ leads to increased likelihood:

$$\max_{\overline{\lambda}}\left[Q\left(\lambda,\overline{\lambda}\right)\right]\Rightarrow P\left(X|\overline{\lambda}\right)\geq P\left(X|\lambda\right) \tag{45}$$

For details

Please look back at our slides on EM!!!

We can rewrite a_{ij}

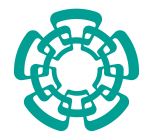
Thus

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i, q_{t}^{(d)} = j | X^{(d)}, \lambda^{n}\right)}{\sum_{d=1}^{D}\sum_{t=2}^{T} P\left(q_{t-1}^{(d)} = i | X^{(d)}, \lambda^{n}\right)}$$

$$= \frac{\sum_{d=1}^{D}\sum_{t=2}^{T} \xi_{q_{t-1}^{(d)}}(i,j)}{\sum_{d=1}^{D}\sum_{t=2}^{T} \gamma_{q_{t-1}^{(d)}}(i)}$$

Remark: This can be seen as

- $\frac{\text{expected number of transitions from state } i \text{ to } j}{\text{expected number of transitions made from state } i}$



Now, the terms $b_i(k)$

We have that

$$\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial b_i(k)} = \frac{\partial}{\partial b_i(k)} \left(\sum_{\mathbf{\Omega}} \sum_{d=1}^{D} \sum_{t=1}^{T} \log b_{q_t^{(d)}} \left(x_t^{(d)} \right) P(\mathbf{\Omega}, \mathcal{X} | \lambda^n) \right) - \lambda_{b_i}$$

$$= \frac{\partial}{\partial b_i(k)} i, \left(\sum_{h=1}^{N} \sum_{d=1}^{D} \sum_{t=1}^{T} \log b_h \left(x_t^{(d)} \right) P\left(q_t^{(d)} = h, \mathcal{X} | \lambda^n \right) \right) - \lambda_{b_i}$$

Now, the terms $b_i(k)$

We have that

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial b_i(k)} &= \frac{\partial}{\partial b_i(k)} \left(\sum_{\mathbf{\Omega}} \sum_{d=1}^{D} \sum_{t=1}^{T} \log b_{q_t^{(d)}} \left(x_t^{(d)} \right) P(\mathbf{\Omega}, \mathcal{X} | \lambda^n) \right) - \lambda_{b_i} \\ &= \frac{\partial}{\partial b_i(k)} i, \left(\sum_{h=1}^{N} \sum_{d=1}^{D} \sum_{t=1}^{T} \log b_h \left(x_t^{(d)} \right) P\left(q_t^{(d)} = h, \mathcal{X} | \lambda^n \right) \right) - \lambda_{b_i}\end{aligned}$$

Now, we have a problem

The term $b_i(k)$ can be different than term $b_i\left(x_t^{(d)}\right)$

We can use this to fix our problem

$$I\left(x_t^{(d)} = k\right) = \begin{cases} 1 & \text{when } x_t^{(d)} == k \\ 0 & \text{Otherwise} \end{cases} \tag{46}$$

Now, the terms $b_i(k)$

We have that

$$\begin{aligned}\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial b_i(k)} &= \frac{\partial}{\partial b_i(k)} \left(\sum_{\mathbf{\Omega}} \sum_{d=1}^{D} \sum_{t=1}^{T} \log b_{q_t^{(d)}} \left(x_t^{(d)} \right) P(\mathbf{\Omega}, \mathcal{X} | \lambda^n) \right) - \lambda_{b_i} \\ &= \frac{\partial}{\partial b_i(k)} i, \left(\sum_{h=1}^{N} \sum_{d=1}^{D} \sum_{t=1}^{T} \log b_h \left(x_t^{(d)} \right) P\left(q_t^{(d)} = h, \mathcal{X} | \lambda^n \right) \right) - \lambda_{b_i}\end{aligned}$$

Now, we have a problem

The term $b_i(k)$ can be different than term $b_i\left(x_t^{(d)}\right)$

We can use this to fix our problem

$$I\left(x_t^{(d)} = k\right) = \begin{cases} 1 & \text{when } x_t^{(d)} == k \\ 0 & \text{Otherwise} \end{cases} \tag{46}$$

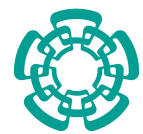
Thus, we have that

Using our previous idea

$$\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial b_i(k)} = \sum_{d=1}^{D}\sum_{t=1}^{T} \frac{P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}{b_i(k)} - \lambda_{b_i}$$

In addition

$$\frac{\partial \hat{L}(\lambda, \lambda^n)}{\partial \lambda_{b_i}} = -\left(\sum_{k=1}^{M} b_i(k) - 1\right) \tag{47}$$



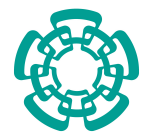
Thus, we have that

Using our previous idea

$$\frac{\partial \hat{L}\left(\lambda, \lambda^{n}\right)}{\partial b_{i}\left(k\right)} = \sum_{d=1}^{D}\sum_{t=1}^{T}\frac{P\left(q_{t}^{(d)} = i, \mathcal{X}|\lambda^{n}\right) I\left(x_{t}^{(d)} = k\right)}{b_{i}\left(k\right)} - \lambda_{b_{i}}$$

In addition

$$\frac{\partial \hat{L}\left(\lambda, \lambda^{n}\right)}{\partial \lambda_{b_{i}}} = -\left(\sum_{k=1}^{M} b_{i}\left(k\right) - 1\right) \tag{47}$$



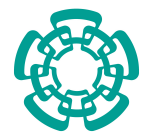
Thus

We have that

$$b_i(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}{\lambda_{b_i}} \tag{48}$$

Thus

$$\lambda_{b_i} = \sum_{k=1}^{M}\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right) \tag{49}$$



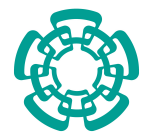
Thus

We have that

$$b_i(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}{\lambda_{b_i}} \tag{48}$$

Thus

$$\lambda_{b_i} = \sum_{k=1}^{M}\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right) \tag{49}$$



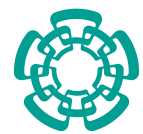
We have then

The following result

$$b_i(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}{\sum_{k=1}^{M}\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}$$

$$= \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = j\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right)}$$

$$= \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i|X^{(d)}, \lambda^n\right) I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i|X^{(d)}, \lambda^n\right)}$$



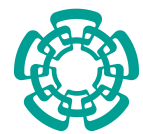
We have then

The following result

$$b_i(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}{\sum_{k=1}^{M}\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}$$

$$= \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = j\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right)}$$

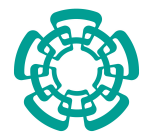
$$= \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i|X^{(d)}, \lambda^n\right) I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i|X^{(d)}, \lambda^n\right)}$$



We have then

The following result

$$
\begin{aligned}
b_i(k) &= \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)}{\sum_{k=1}^{M}\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = k\right)} \\
&= \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right) I\left(x_t^{(d)} = j\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i, \mathcal{X}|\lambda^n\right)} \\
&= \frac{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i|X^{(d)}, \lambda^n\right) I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T} P\left(q_t^{(d)} = i|X^{(d)}, \lambda^n\right)}
\end{aligned}
$$



Using Rabiner's Notation

We have that

$$b_i(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)\, I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)} \tag{50}$$

Which can be seen as

$$\frac{\text{The expected number of times in state } i \text{ and observing symbol } x_k}{\text{The expected number of times in state } i}$$

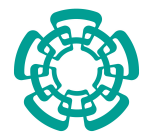
Using Rabiner's Notation

We have that

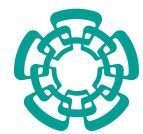
$$b_i(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)\, I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)} \tag{50}$$

Which can be seen as

$$\frac{\text{The expected number of times in state } i \text{ and observing symbol } x_k}{\text{The expected number of times in state } i}$$



Outline



Thus, the re-estimation

First

$$\pi_i^{(n+1)} = \frac{1}{D}\sum_{d=1}^{D}\gamma_{q_1^{(d)}}(i) \tag{51}$$

Second

$$a_{ij}^{(n+1)} = \frac{\sum_{d=1}^{D}\sum_{t=2}^{T}\xi_{q_{t-1}^{(d)}}(i,j)}{\sum_{d=1}^{D}\sum_{t=2}^{T}\gamma_{q_{t-1}^{(d)}}(i)} \tag{52}$$

Third

$$b_i^{(n+1)}(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)\, I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)} \tag{53}$$

Thus, the re-estimation

First

$$\pi_i^{(n+1)} = \frac{1}{D}\sum_{d=1}^{D}\gamma_{q_1^{(d)}}(i) \tag{51}$$

Second

$$a_{ij}^{(n+1)}\frac{\sum_{d=1}^{D}\sum_{t=2}^{T}\xi_{q_{t-1}^{(d)}}(i,j)}{\sum_{d=1}^{D}\sum_{t=2}^{T}\gamma_{q_{t-1}^{(d)}}(i)} \tag{52}$$

Third

$$b_i^{(n+1)}(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)\, I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)} \tag{53}$$

Thus, the re-estimation

First

$$\pi_i^{(n+1)} = \frac{1}{D}\sum_{d=1}^{D}\gamma_{q_1^{(d)}}(i) \tag{51}$$

Second

$$a_{ij}^{(n+1)}\frac{\sum_{d=1}^{D}\sum_{t=2}^{T}\xi_{q_{t-1}^{(d)}}(i,j)}{\sum_{d=1}^{D}\sum_{t=2}^{T}\gamma_{q_{t-1}^{(d)}}(i)} \tag{52}$$

Third

$$b_i^{(n+1)}(k) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)\, I\left(x_t^{(d)} = k\right)}{\sum_{d=1}^{D}\sum_{t=1}^{T}\gamma_{q_t^{(d)}}(i)} \tag{53}$$

Something Nice

An important aspect of the re-estimation procedure is that

$$\sum_{i=1}^{N} \pi_i^{(n+1)} = 1$$

$$\sum_{j=1}^{N} a_{ij}^{(n+1)} = 1 \; 1 \leq i \leq N$$

$$\sum_{k=1}^{M} b_j^{(n+1)}(k) = 1 \; 1 \leq j \leq N$$

They are

Automatically it is satisfied at each iteration.

Something Nice

An important aspect of the re-estimation procedure is that

$$\sum_{i=1}^{N} \pi_i^{(n+1)} = 1$$

$$\sum_{j=1}^{N} a_{ij}^{(n+1)} = 1 \; 1 \leq i \leq N$$

$$\sum_{k=1}^{M} b_j^{(n+1)}(k) = 1 \; 1 \leq j \leq N$$

They are

Automatically it is satisfied at each iteration.

Outline

Once these probabilities are ready

We can use Naïve Bayes

$$P(\Omega_i|X) = \frac{P(\Omega_i, X)}{P(X)} > \frac{P(\Omega_j, X)}{P(X)} = P(\Omega_j|X) \ \forall i \neq j \tag{54}$$

Then

The rule looks like

$$P(\Omega_i|X)\,P(\Omega_i) > P(\Omega_j|X)\,P(\Omega_j) \ \forall i \neq j \tag{55}$$

Once these probabilities are ready

We can use Naïve Bayes

$$P(\Omega_i|X) = \frac{P(\Omega_i, X)}{P(X)} > \frac{P(\Omega_j, X)}{P(X)} = P(\Omega_j|X) \ \forall i \neq j \tag{54}$$

Then

The rule looks like

$$P(\Omega_i|X)\,P(\Omega_i) > P(\Omega_j|X)\,P(\Omega_j) \ \forall i \neq j \tag{55}$$