Introduction to hidden Markov models

Alexis Huet

8 April 2013

Outline

Outline

1. Introduction
2. Likelihood of the observations
   - Brute-force
   - Forward decomposition
3. Computation of the best hidden sequence
   - Definition and method (Viterbi algorithm)
4. Optimization of the model parameters
   - Brute-force
   - Hard Expectation-Maximization algorithm
5. Real applications of hidden Markov models

Markov chains

Let E a finite state space with N elements.

Definition

A sequence of random variables $(X_k)_{k\in\mathbb{N}}$ taking values in E is a Markov chain if for all $n \geq 1$ and $x_1, \ldots x_n \in E$:

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}).$$

Definition

A Markov chain $(X_k)_{k\in\mathbb{N}}$ is said homogeneous if for all $i, j \in E$ and $n \geq 1$:

$$P(X_n = j | X_{n-1} = i) = P(X_1 = j | X_0 = i).$$

Markov chains

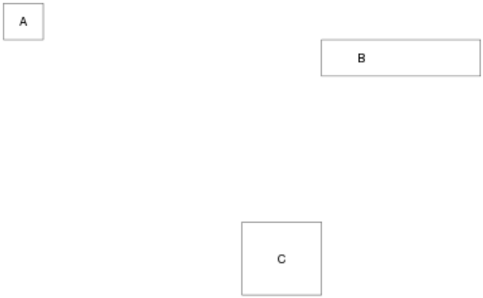
In the sequel, $(X_k)_{k\in\mathbb{N}}$ is an homogeneous Markov chain taking values in $E = (e_1, \dots, e_N)$.

Property

$(X_k)_{k\in\mathbb{N}}$ is characterized by :

- *the row vector π defined for all i by : $\pi(i) = P(X_0 = e_i)$.*
- *the transition matrix M defined for all i, j by : $M(i,j) = P(X_1 = j|X_0 = i)$*

Example



Example

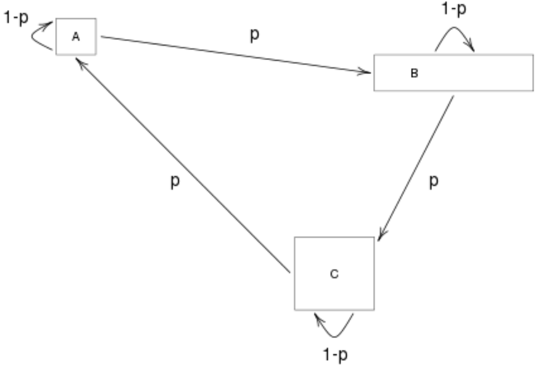
We take the following initial condition :

$$\begin{array}{cc} & \begin{array}{ccc} A & B & C \end{array} \\ \pi = & \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \end{array}$$

and this transition matrix :

$$M = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} \quad A \quad & \quad B \quad & \quad C \quad \end{array} \\ \begin{pmatrix} 1-p & p & 0 \\ 0 & 1-p & p \\ p & 0 & 1-p \end{pmatrix} \end{array}.$$

Example



Example

We obtain a sequence in the form of :

$$X_0 \longrightarrow X_1 \longrightarrow X_2 \longrightarrow \dots \longrightarrow X_{m-1}.$$

For $p = 0.4$, a length $m = 100$ and a randomness ω, we get the following sequence :
C C C A B B B C A B B B B B C C A A A B C C C A A B B C C C C C A A A A B B B B B C C C C A A A A A A A B C A B B B B B B B B B B C C A A B C A A B B B C A A B B C C A B B B B B B C C C A A A A A A B.

Hidden Markov models

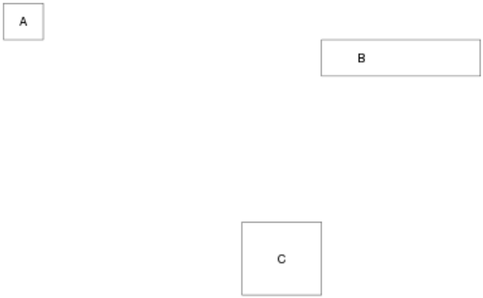
Definition

$(X_k, Y_k)_{k\in 0:m-1}$ is a hidden Markov model if : $(X_k)_{k\in 0:m-1}$ is a Markov chain, $(Y_k)_{k\in 0:m-1}$ are independent conditionally to $(X_k)_{k\in 0:m-1}$ and for all k, Y_k depends only on X_k.

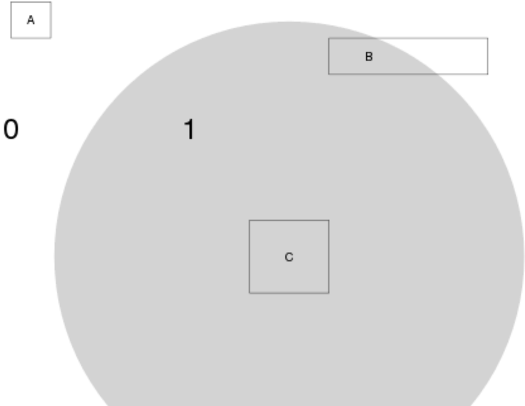
Schematically, we have :

$$\begin{array}{ccccccccc} X_0 & \longrightarrow & X_1 & \longrightarrow & X_2 & \longrightarrow & \dots & \longrightarrow & X_{m-1} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ Y_0 & & Y_1 & & Y_2 & & \dots & & Y_{m-1} \end{array}.$$

Example



Example



Example

For all site k, the transition to Y_k conditionally to X_k is given by the matrix :

$$N = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \begin{pmatrix} 1 & 0 \\ 1-q & q \\ 0 & 1 \end{pmatrix} \end{array}.$$

For $p = 0.4$ and $q = 0.7$, we get the sequence :

$$\begin{array}{ccccccccccc} C & \longrightarrow & C & \longrightarrow & C & \longrightarrow & A & \longrightarrow & \dots & \longrightarrow & B \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ 1 & & 1 & & 1 & & 0 & & \dots & & 1 \end{array}.$$

Example

```
C C C A B B B C A B B B B B C C A A A B C C C A A B B
1 1 1 0 1 0 0 1 0 0 1 1 0 0 1 1 0 0 0 1 1 1 1 0 0 1 1

C C C C C A A A A B B B B B C C C C A A A A A A A B C
1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1

A B B B B B B B B B B C C A A B C A A B B B C A A B B
0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 1 1 0 0 1 0 0 1 0 0 0 1

C C A B B B B B B C C C A A A A A A B
1 1 0 0 1 0 1 0 0 1 1 1 0 0 0 0 0 0 1.
```

Example

1 1 1 0 1 0 0 1 0 0 1 1 0 0 1 1 0 0 0 1 1 1 1 0 0 1 1

1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1

0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 1 1 0 0 1 0 0 1 0 0 0 1

1 1 0 0 1 0 1 0 0 1 1 1 0 0 0 0 0 0 1.

Issues

Now, the hidden chain (x_k) is unknown and we only have the observations (y_k). We want :

- knowing the model, to compute the likelihood of the observations.
- knowing the model, to fit the hidden sequence with the highest likelihood.
- to estimate the parameters of the model.

Outline

Likelihood computation

From now, we assume that we know the observed values $y_{0:m-1}$. Moreover, the model is fixed here (initial distribution and transition matrix).

Aim

Compute $p(y_{0:m-1})$ likelihood of the observed values.

As the model is fixed, we can calculate, for all x, x', y :

$$p(Y_k = y | X_k = x) \text{ and } p(X_{k+1} = x' | X_k = x)$$

written in the next slides :

$$p(y|x) \text{ and } p(x'|x).$$

Likelihood computation : brute-force

For an hidden sequence $x_{0:m-1}$, we have :

$$p(x_{0:m-1}, y_{0:m-1}) = \pi(x_0) \prod_{k=0}^{m-1} p(y_k | x_k) \prod_{k=0}^{m-2} p(x_{k+1} | x_k).$$

Thus :

$$p(y_{0:m-1}) = \sum_{x_{0:m-1}} \pi(x_0) \prod_{k=0}^{m-1} p(y_k | x_k) \prod_{k=0}^{m-2} p(x_{k+1} | x_k).$$

The sum is on the $|E|^m$ elements. Cannot be used when m increases.

Likelihood computation : forward decomposition

The Markovian structure is used. To compute, for all $k \in 0 : m-1$, $i \in E$:

$$\alpha_k(i) = p(Y_{0:k} = y_{0:k}, X_k = i).$$

We write :

$$\begin{aligned}
\alpha_{k+1}(j) =& p(Y_{0:k+1} = y_{0:k+1}, X_{k+1} = j) \\
=& p(y_{k+1} | y_{0:k}, X_{k+1} = j) p(y_{0:k}, X_{k+1} = j) \\
=& p(y_{k+1} | X_{k+1} = j) \sum_i p(y_{0:k}, X_k = i, X_{k+1} = j) \\
=& p(y_{k+1} | X_{k+1} = j) \sum_i p(X_{k+1} = j | y_{0:k}, X_k = i) P(y_{0:k}, X_k = i) \\
=& p(y_{k+1} | X_{k+1} = j) \sum_i p(X_{k+1} = j | X_k = i) \alpha_k(i).
\end{aligned}$$

Likelihood computation : forward decomposition

With :

$$\alpha_k(i) = p(Y_{0:k} = y_{0:k}, X_k = i).$$

- Initialization : $\alpha_0(i) = \pi(i)p(y_0|X_0 = i)$.
- Induction :
  $\alpha_{k+1}(j) = p(y_{k+1}|X_{k+1} = j)\sum_i p(X_{k+1} = j|X_k = i)\alpha_k(i)$.
- Likelihood computation :

$$p(y_{0:m-1}) = \sum_i \alpha_{m-1}(i).$$

Complexity : $|E|^2 m$, linear with the length of the sequence.

Example

With the previous example, with $p = 0.4$, $q = 0.7$ and the observed sequence 1 1 1 0 1 0 0 1 0 ... 0 0 0 0 0 1, we get :
For $j \in \{A, B, C\}$,

$$\alpha_0(j) = \pi(j)p(y_0|X_0 = j) = \mathbf{1}_{j=C}.$$

$$\begin{aligned}
\alpha_1(j) &= p(y_1|X_1 = j)\sum_i p(X_1 = j|X_0 = i)\alpha_0(i) \\
&= p(1|X_1 = j)p(X_1 = j|X_0 = C) \\
&= (1-p)\mathbf{1}_{j=C} \\
&= 0.6 \times \mathbf{1}_{j=C}.
\end{aligned}$$

etc.

Example

Sequence : 1 1 1 0 1 0 0 1 0 . . . 0 0 0 0 0 1

$$\begin{array}{c} \\ \alpha_0 = \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \end{array}, \quad \alpha_1 = (0, 0, 0.6), \quad \alpha_2 = (0, 0, 0.36),$$

$$\alpha_3 = (0.144, 0, 0),$$

$$\alpha_4 = (0, 0.04, 0), \quad \alpha_5 = (0, 0.007, 0), \quad \alpha_6 = (0, 0.001, 0),$$

$$\alpha_7 = (0, 5.49e{-}04, 5.23e{-}04), \quad \alpha_8 = (2.09e{-}04, 9.88e{-}05, 0), \ldots,$$

$$\alpha_{m-1} = (0, 1.00e - 30, 2.86e - 31).$$

Thus $p(y_{0:m-1}|p = 0.4, q = 0.7) = 1.29e - 30$.

Outline

Introducing the problem

We still have the observed values $y_{0:m-1}$, the model is fixed. The aim is to seek the best sequence $x_{0:m-1}$ in the following sense, knowing the observed values.

Aim

Compute $\arg\max_{x_{0:m-1}} p(x_{0:m-1}, y_{0:m-1})$.

To do that, we use the Viterbi algorithm.

Idea of the algorithm

Out aim is to compute
$(x_0^*, \ldots, x_{m-1}^*) = \arg\max_{x_{0:m-1}} p(x_{0:m-1}, y_{0:m-1})$.
We assume that we have $x_{k+1}^*, \ldots, x_{m-1}^*$. Then :

$$
\begin{aligned}
(x_0^*, \ldots, x_k^*) &= \arg\max_{x_{0:k}} p(x_{0:k}, x_{k+1:m-1}^*, y_{0:m-1}) \\
&= \arg\max_{x_{0:k}} p(x_{0:k}, y_{0:k}) p(x_{k+1}^* | x_k) p(x_{k+2:m-1}^*, y_{k+1:m-1} | x_{k+1}^*) \\
&= \arg\max_{x_{0:k}} p(x_{0:k}, y_{0:k}) p(x_{k+1}^* | x_k).
\end{aligned}
$$

Thus : $x_k^* = \arg\max_{x_k} \underbrace{\underbrace{\max_{x_{0:k-1}} [p(x_{0:k}, y_{0:k})]}_{\delta_k(x_k)} p(x_{k+1}^* | x_k)}_{\psi_{k+1}(x_{k+1}^*)}$.

Viterbi algorithm

For all site k, for all hidden state $i \in E$, we let :

$$\delta_k(i) = \max_{x_{0:k-1}} p(y_{0:k}, x_{0:k-1}, X_k = i).$$

We check for $j \in E$ (same method as the forward process) :

$$\delta_{k+1}(j) = p(y_{k+1}|X_{k+1} = j) \max_i [\delta_k(i) p(X_{k+1} = j|X_k = i)].$$

Finally :

- Initialization : $\delta_0(i) = \pi(i) p(y_0|X_0 = i)$.
- Induction : $\delta_{k+1}(j)$ according to the above formula.
- Return initialization : $x^*_{m-1} = \arg\max_{x_{m-1}} \delta_{m-1}(x_{m-1})$.
- Return : $x^*_k = \psi_{k+1}(x^*_{k+1})$ with :

$$\psi_{k+1}(j) = \arg\max_{x_k} \delta_k(x_k) p(X_{k+1} = j|x_k).$$

Example

C C C A B B B C A B B B B B C C A A A B C C C A A B B
C C C A B B B C A A B C A A B C A A A B C C C A A B C
1 1 1 0 1 0 0 1 0 0 1 1 0 0 1 1 0 0 0 1 1 1 1 0 0 1 1

C C C C C A A A A B B B B B C C C C A A A A A A A B C
C C C C C A A A A A A A B C C C C C A A A A A A A B C
1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1

A B B B B B B B B B B C C A A B C A A B B B C A A B B
A A A A B B C A B B B C C A A B C A A B B B C A A A B
0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 1 1 0 0 1 0 0 1 0 0 0 1

C C A B B B B B B C C C A A A A A A B
C C A A B B C A A B C C A A A A A A B
1 1 0 0 1 0 1 0 0 1 1 1 0 0 0 0 0 0 1.

Outline

Introducing the problem

We know the observed values $y_{0:m-1}$. The model now depends on parameters $\theta \in \Theta$.

Aim

Compute $\arg\max_\theta p(y_{0:m-1}|\theta)$ most probable parameters of the model.

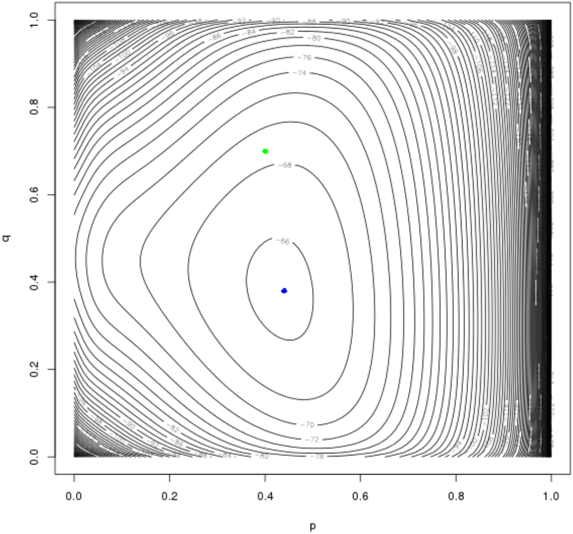
Two methods are set out here :

- Use the first part of this talk and compute $p(y_{0:m-1}|\theta)$ for all parameters.
- Use the second part and recursively update the parameters, depending of the best hidden sequence found.

Example

We take again the sequence :
1 1 1 0 1 0 0 1 0 0 1 1 0 0 . . . 0 0 0 0 0 1.

We seek a parameter $\theta = (p, q) \in [0, 1] \times [0, 1]$. We calculate $\log p(y_{0:m-1}|p, q)$ with a step of 0.01, and then take the maximum.


q
p
0.0
0.2
0.4
0.6
0.8
1.0
-66
-68
-70
-72
-74
-76
-78
-80

Hard Expectation-Maximization algorithm

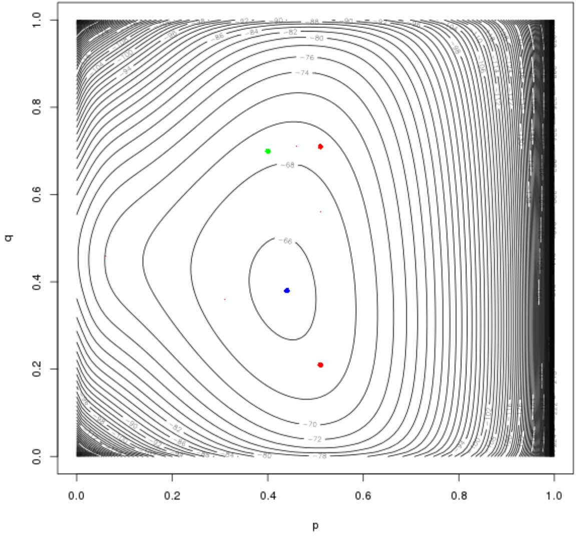
The observed values $y = y_{0:m-1}$ are known. We let $\theta_0 \in \Theta$ some initial parameters.

For $i \geq 0$:

- Compute x_i with the Viterbi algorithm, for y and θ_i.
- Maximize the couple (x_i, y) over the set of the parameters :

$$\theta_{i+1} = \arg\max_{\theta} p(x_i, y|\theta).$$

The estimation of the parameters is the last θ_i computed.


q
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
p
−76
−74
−68
−66
−70
−72

Outline

Phylogenetic analysis

- Observations : DNA sequences of several species at the leafs of a tree graph.
- Hidden states : all DNA sequences from the common ancestry sequence to the present time.
- Parameters : mutation parameters, lengths of the tree branches.

Voice recognition system

- Observations : a word is pronounced, cut every 15ms.
- Hidden states : phonemes that led to this pronounced word.
- Parameters : the set of all dictionary words.

Path tracking

- Observations : noisy position.
- Hidden states : real position.
- Parameters : behavior of the moving body.

Thank you for your attention !