

SI221 : Bases de l'Apprentissage

Modèles de Markov Cachés

Mars 2022

Laurence Likforman-Sulem
Telecom Paris/IP-Paris
likforman@telecom-paristech.fr



1

Plan

- Chaînes de Markov
 - modèles stochastiques
 - Paramètres
 - Génération de séquences états
- Modèles de Markov Cachés
 - discrets/continus
 - apprentissage
 - décodage

Laurence Likforman-Telecom ParisTech

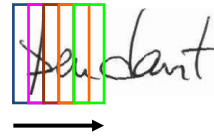
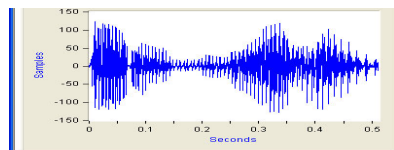
2

2

applications

■ HMMs

- Reconnaissance de la parole (speech recognition)
- Reconnaissance de l'écriture (handwriting recognition)
- Reconnaissance d'objets, de visages dans les vidéos
- Traitement de la langue-Natural Language Processing-NLP: étiquetage grammatical, correction orthographique



THE → TGE



Laurence Likforman-Telecom ParisTech

3

3

Modèle stochastique

- processus aléatoire à temps discret
 - ensemble de variables aléatoires q_1, q_2, \dots, q_T
 - indexées aux instants entiers $t=1, 2, \dots, T$
- notation
 - q_t : variable aléatoire d'état observé au temps t
 - notée $q(t)$ ou q_t
 - $q(t)$ prend ses valeurs dans espace fini d'états S
 $S=\{1, 2, \dots, Q\}$
 - $P(q_t=i)$: probabilité d'observer l'état i au temps t

exemple état: pollution (indice), météo: beau, pluie, nuageux, NLP: fonction des mots d'un texte (verbe, nom, pronom;....)

Laurence Likforman-Telecom ParisTech

4

4

Modèle stochastique

- évolution du processus
 - état initial q_1
 - suite (chaîne) de transitions entre états
 - $q_1 \rightarrow q_2 \dots \rightarrow q_t \quad t \leq T$
- calcul probabilité d'une séquence d'états
$$\begin{aligned} P(q_1, q_2, \dots, q_T) &= P(q_T | q_1, q_2, \dots, q_{T-1}) P(q_1, q_2, \dots, q_{T-1}) \\ &= P(q_T | q_1, q_2, \dots, q_{T-1}) P(q_{T-1} | q_1, q_2, \dots, q_{T-2}) P(q_1, q_2, \dots, q_{T-2}) \\ &= P(q_1) P(q_2 | q_1) P(q_3 | q_1, q_2) \dots P(q_T | q_1, q_2, \dots, q_{T-1}) \end{aligned}$$
- modèle: connaître la probabilité de chaque transition+proba initiale $P(q_1)$

5

5

Chaîne de Markov à temps discret

- propriété de Markov d'ordre k : dépendance limitée
 - $P(q_t | q_1, q_2, \dots, q_{t-1}) = P(q_t | q_{t-k}, \dots, q_{t-1})$
 - $k=1$ ou 2 en pratique
- cas $k=1$ (bi-grams)
 - $P(q_t | q_1, q_2, \dots, q_{t-1}) = P(q_t | q_{t-1})$
 - $P(q_1, q_2, \dots, q_T) = P(q_1) P(q_2 | q_1) P(q_3 | q_2) \dots P(q_T | q_{T-1})$
 - \rightarrow probabilités de transition entre états

Laurence Likforman-Telecom ParisTech

6

6

Chaîne de Markov stationnaire

- probabilités de transition ne dépendent pas du temps
 - $P(q_t = j \mid q_{t-1} = i) = P(q_{t+k} = j \mid q_{t+k-1} = i) = a_{ij}$
 - a_{ij} = probabilité de passer de l'état i à l'état j
- définition: modèle d'une chaîne de Markov stationnaire
 - matrice des probabilités de transitions
 - $A = [a_{ij}] \quad i=1, \dots, Q, j=1, \dots, Q$
 - vecteur des probabilités initiales
 - $\Pi = [\pi_i] \quad i=1, \dots, Q$
 - $\pi_i = P(q_1 = i)$
- contraintes : $0 \leq \pi_i \leq 1 \quad 0 \leq a_{ij} \leq 1$

$$\sum_{i=1}^Q \pi_i = 1$$

$$\sum_{j=1}^Q a_{ij} = 1$$

Laurence Likforman-Telecom ParisTech

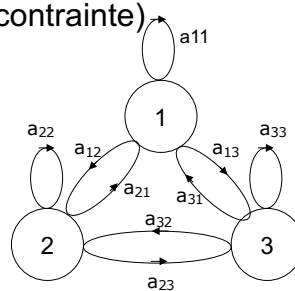
7

7

topologie du modèle: ergodique / gauche droite

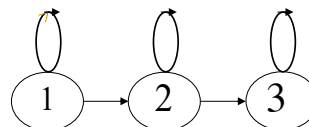
- modèle ergodique (sans contrainte)

$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$



- modèle gauche droite (contrainte: transitions $i \rightarrow j \geq i$)

$$A = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}$$



8

8

Chaîne de Markov stationnaire: mini TD

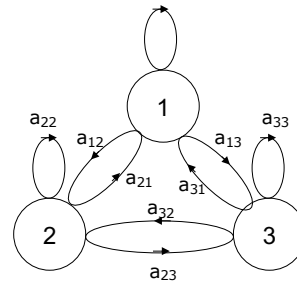
- Soit une chaîne à 3 états
 - 1: pluie (r), 2: nuages (c) , 3: soleil (s)
- on observe $q_1 = s$, quelle est la probabilité d'observer pendant les 7 jours suivants les temps (états)

s s s r r s c s

- $t=1$ $t=2$

- modèle ergodique

$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

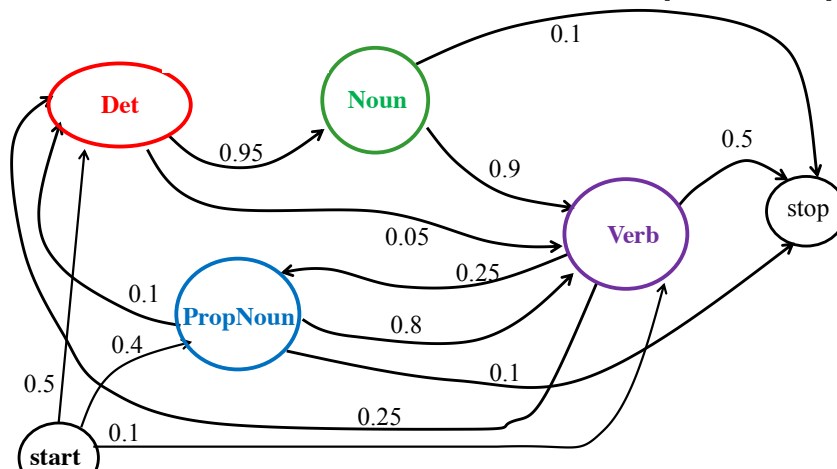


9

9

Mini-TD: POS part-of-speech tagging

[from C. Cardie]



Probabilité de la séquence: Nom propre-verbe -déterminant – Nom ?

10

10

générer une séquence d'états: mini-TD

- on donne
- générer séquence d'états de longueur $T=5$ suivant chaîne de Markov (matrice A)
- $\pi = [0.35 \ 0.65]$ $A = \begin{bmatrix} 0.35 & 0.65 \\ 0.2 & 0.8 \end{bmatrix}$
- on tire les nombres aléatoires suivants:
- $u1 = 0.92$ ($q1$)
- $u2 = 0.31$
- $u3 = 0.1$
- $u4 = 0.4$
- $u5 = 0.01$

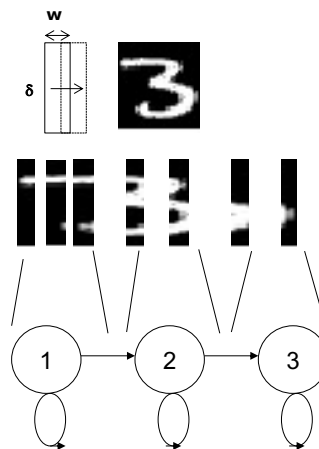
Laurence Likforman-Telecom ParisTech

13

13

Modèles de Markov Cachés

- une classe c de forme
 - modèle λ_c
- combinaison de 2 processus stochastiques
 - un observé
 - un caché
- on n'observe pas la séquence d'états
 $q = q_1 q_2 \dots q_T$
- on observe la séquence d'observations
 $O = O_1 O_2 \dots O_T$
- les observations sont générées (émises) par les états



Laurence Likforman-Telecom ParisTech

14

14

HMMs discrets

- ensemble de Q états discrets $\{1, 2, \dots, Q\}$
- ensemble de N symboles discrets
 - $\{s_1, s_2, s_3, \dots, s_N\} \rightarrow \{1, 2, 3, \dots, N\}$

- on observe $o = o_1 o_2 o_t \dots o_T$

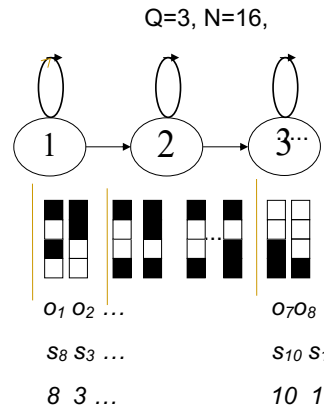
- $o = s_8 s_3 s_{13} s_6 s_8 s_5 s_{10} s_1$

- $o = 8 \ 3 \ 13 \ 6 \ 8 \ 5 \ 10 \ 1$

- q correspond à séquence d'états (cachés)

- $q = q_1 q_2 q_t \dots q_T$

- $q = 1 \ 1 \ 2 \ 2 \ 2 \ 2 \ 3 \ 3$



15

15

HMMs discrets

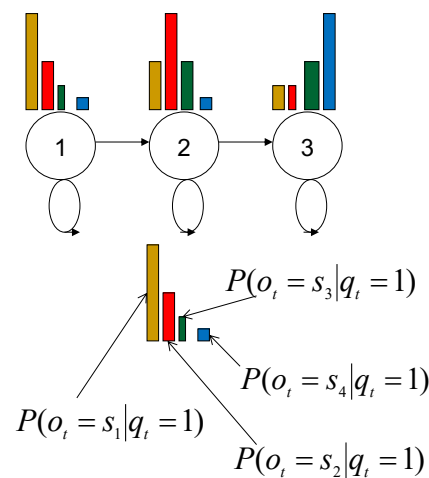
- HMM λ discret est défini par

- π vecteur probabilités initiales
- A : matrice transition
- B : matrice des probabilités discrètes d'observation des symboles (dans les états)

$$\pi = (\pi_1, \pi_2, \dots, \pi_Q) \quad \pi_i = P(q_1 = i)$$

$$A = \{a_{ij}\} = P(q_t = j | q_{t-1} = i)$$

$$B = \{b_{ki}\} = P(o_t = s_k | q_t = i)$$



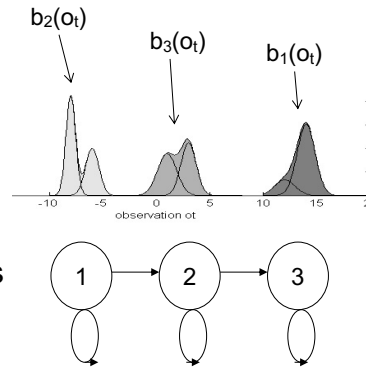
Laurence Likforman-Telecom
ParisTech

16

16

modèles de Markov cachés continus

- HMM λ continu défini par :
- π vecteur de probabilités initiales
- A : matrice de transition entre états
- $b_i(o_t)$: densité de probabilité des observations dans état i , $i=1,..Q$
→ gaussienne ou mélange gaussiennes

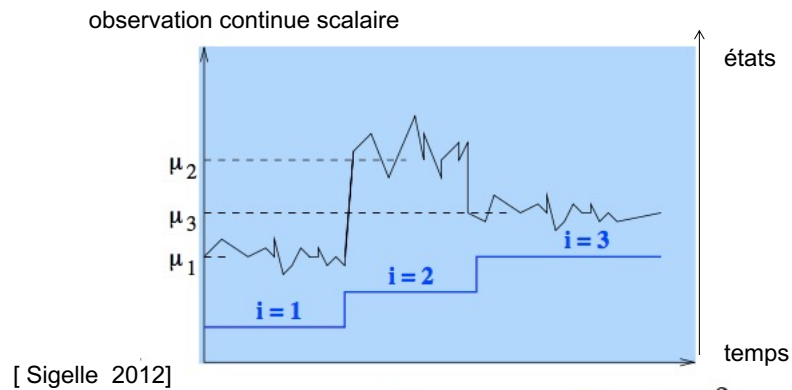


L. Likforman - Telecom ParisTech

17

17

modèle d'observations Gaussien



$$p(o_t / q_t = i, \lambda) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp - \frac{(o_t - \mu_i)^2}{2\sigma_i^2}$$

modèle: inclut μ_i et σ_i , $i=1,2,3$

Laurence Likforman-Telecom ParisTech

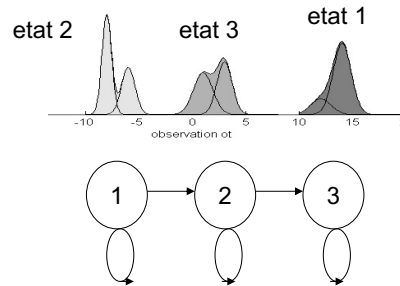
18

18

mélange de gaussiennes

$$b_i(o_t) = \sum_{k=1}^M c_{ik} \mathcal{N}(o_t; \Sigma_{ik}, \mu_{ik}) \quad \forall i = 1, \dots, Q.$$

observations continues (scalaires ou vectorielles)



c_{ik} : poids de la k ème loi gaussienne du mélange de M gaussiennes, associée à l'état i

modèle λ : inclut c_{ik} , μ_{ik} et Σ_{ik} , $i=1, 2, 3$ et $k=1, \dots, M$

L. Likforman - Telecom ParisTech

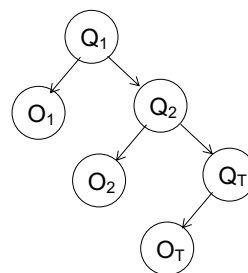
19

19

hypothèses fondamentales

- indépendance des observations conditionnellement aux états

$$P(o_1, \dots, o_T | q_1, \dots, q_T, \lambda) = \prod_{t=1}^T P(o_t | q_t, \lambda)$$



- chaîne de Markov stationnaire (transitions entre états)

$$P(q_1, q_2, \dots, q_T) = P(q_1) P(q_2 | q_1) P(q_3 | q_2) \dots P(q_T | q_{T-1})$$

$\underbrace{\hspace{1.5cm}}_{a_{q_2 q_3}}$

Laurence Likforman-Telecom ParisTech

20

20

hypothèses fondamentales

- probabilité jointe pour une séquence d'observations et un chemin d'états

$$\begin{aligned} P(o_1, \dots, o_T, q_1, \dots, q_T | \lambda) &= \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^T a_{q_{t-1}, q_t} P(o_t | q_t, \lambda) \\ &= \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^T a_{q_{t-1}, q_t} b_{q_t}(o_t) \\ &= P(o_1, \dots, o_T | q_1, \dots, q_T, \lambda) P(q_1, \dots, q_T) \end{aligned}$$

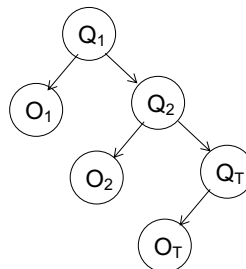
Laurence Likforman-Telecom ParisTech

21

21

HMM / réseau bayésien

- un HMM est un cas particulier de réseau Bayésien
- les variables d'observations sont indépendantes connaissant leur variable parent (état)

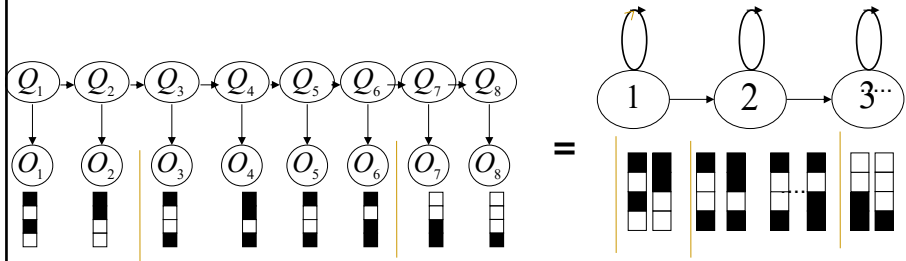


Laurence Likforman-Telecom
ParisTech

22

22

HMM= cas particulier de réseau bayésien dynamique



- HMM: Hidden Markov Model
- RBD: réseau Bayésien Dynamique de type arbre
- 1 variable état + 1 variable observation à chaque t

$(Q_t)_{1 \leq t \leq T}$: state variable (hidden)

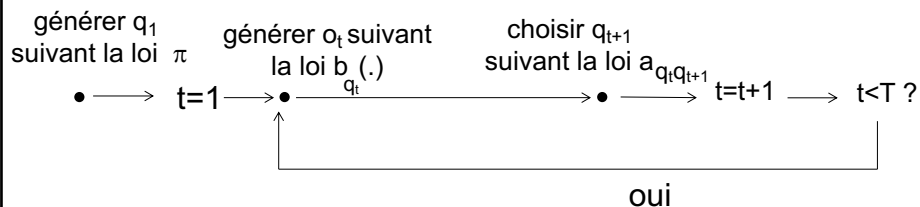
$(O_t)_{1 \leq t \leq T}$: observation variable generated by state variable

23

23

générer une séquence d'observations

- générer la séquence d'états q_1, \dots, q_T , puis générer la séquence observations à partir de chaque état
- ou générer q_1 puis o_1 ($q_1 \rightarrow o_1$); générer q_2 à partir de q_1 ($q_1 \rightarrow q_2$), puis o_2 ($q_2 \rightarrow o_2$), etc...



L. Likforman - Telecom ParisTech

24

24

étape 2 : générer des observations (discretes)

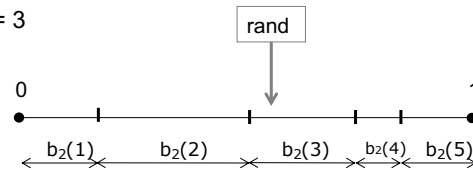
- séquence états

- $q_1=1; q_2=1; q_3=1; q_4=2; q_5=2; q_6=3; \dots$

- générer l'observation à $t=4$

- $q_4=2;$

- $\rightarrow o_4=3$



- $N=5$ (nombre de symboles)

- $Q=4$ (nombre d'états)

		états				
B=	[0.3	0.	0.1	0.2	observations
		0.1	0.7	0.2	0.1	
		0.5	0.1	0.2	0.1	
		0.1	0.1	0.3	0.3	
		0	0.1	0.2	0.3	

25

HMM pour la reconnaissance des formes

- chaque classe m est modélisée par un modèle λ_m

- calcul de la vraisemblance du modèle λ_m pour une séquence d'observations $o=o_1, \dots, o_T$ extraite d'une forme

$$P(o_1, \dots, o_t, \dots, o_T | \lambda_m)$$

- attribution de la forme à la classe \hat{m} telle que :

$$\hat{m} = \arg \max_m P(o_1, \dots, o_t, \dots, o_T | \lambda_m)$$

26

26

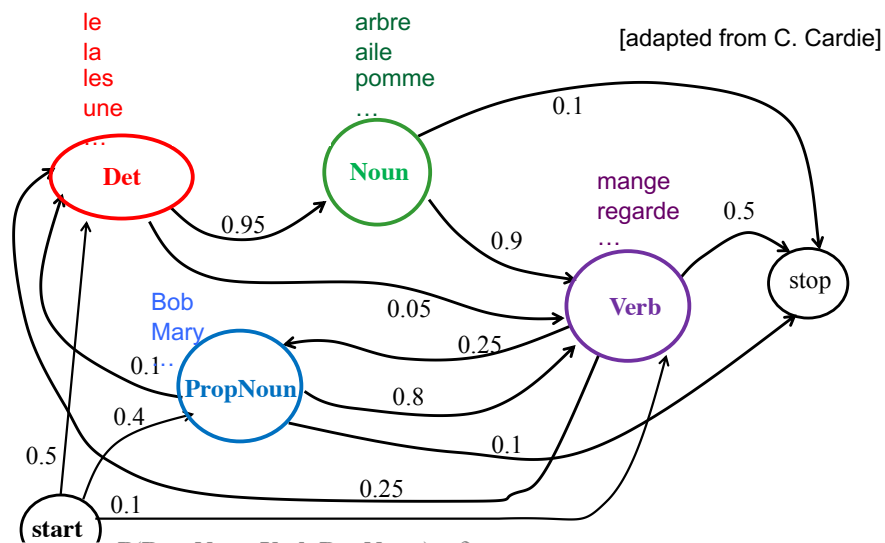
HMM pour étiquetage morpho-syntaxique

- observations: mots
- séquence d'observation : suite de mots
- états cachés: Nom, pronom, verbe, etc....
- modèle
 - probabilités de transitions entre éléments grammaticaux, bi-grams (tags)
 - probabilités d'observer les mots pour un élément grammatical donné (tag)
 - $P(\text{« le »} \mid \text{verbe})$, $P(\text{« le »} \mid \text{pronom})$ etc....

27

27

Mini-TD: POS part-of-speech tagging



Probabilité de la séquence: Nom propre-verbe -déterminant - Nom) ?

28

28

algorithme de décodage de Viterbi

- calcul de la vraisemblance du modèle λ
- séquence observation $o=o_1, \dots, o_T$

$$P(o | \lambda) = \sum_q P(o, q | \lambda)$$

- au lieu de sommer sur toutes les séquences d'états, recherche de la séquence optimale :

$$\hat{q} = \arg \max_q P(q, o | \lambda)$$

- puis estimer la vraisemblance du modèle λ par :

$$P(o | \lambda) \approx P(o, \hat{q} | \lambda)$$

Laurence Likforman-Telecom
ParisTech

29

29

décodage : algorithme de Viterbi

- $\delta_t(i)$: proba. (jointe) meilleure séquence partielle d'états aboutissant à l'état i au temps t et correspondant à la séquence partielle d'observations $o_1 \dots o_t$.

$$\delta_t(i) = \max_{q_1 q_2 \dots q_{t-1}} P(q_1 q_2 \dots q_t = i, o_1 o_2 \dots o_t | \lambda)$$

- récurrence

$$P(q_1 q_2 \dots q_t = i, q_{t+1} = j, o_1 o_2 \dots o_t o_{t+1} | \lambda)$$

$$= P(o_{t+1}, q_{t+1} = j | o_1 \dots o_t, q_1 \dots q_t = i, \lambda) P(o_1 \dots o_t, q_1 \dots q_t = i | \lambda)$$

$$= P(o_{t+1} | q_{t+1} = j, \lambda) P(q_{t+1} = j | q_t = i, \lambda) P(o_1 \dots o_t, q_1 \dots q_t = i | \lambda)$$

$$\max_i P(q_1 q_2 \dots q_t = i, q_{t+1} = j, o_1 o_2 \dots o_t o_{t+1} | \lambda) = \max_i b_j(o_{t+1}) a_{ij} P(q_1 q_2 \dots q_t = i, o_1 o_2 \dots o_t | \lambda)$$

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

$$P(o, \hat{q}) = \max_j \delta_T(j)$$

Laurence Likforman-Telecom ParisTech

30

30

algorithme de décodage de Viterbi

- 1ere colonne: Initialisation

$$\delta_1(i) = P(q_1 = i, o_1) = b_i(o_1)\pi_i \quad i = 1, \dots, Q$$

- colonnes 2 à T : récursion

$$\delta_{t+1}(j) = b_j(o_{t+1}) \max_i a_{ij} \delta_t(i) \quad t = 1, \dots, T-1, j = 1, \dots, Q$$

$$\varphi_{t+1}(j) = \arg \max_i a_{ij} \delta_t(i) \quad \text{sauvegarde meilleur chemin (état précédent)}$$

- terminaison

$$P(o, \hat{q}) = \max_j \delta_T(j)$$

$$\hat{q}_T = \arg \max_j \delta_T(j)$$

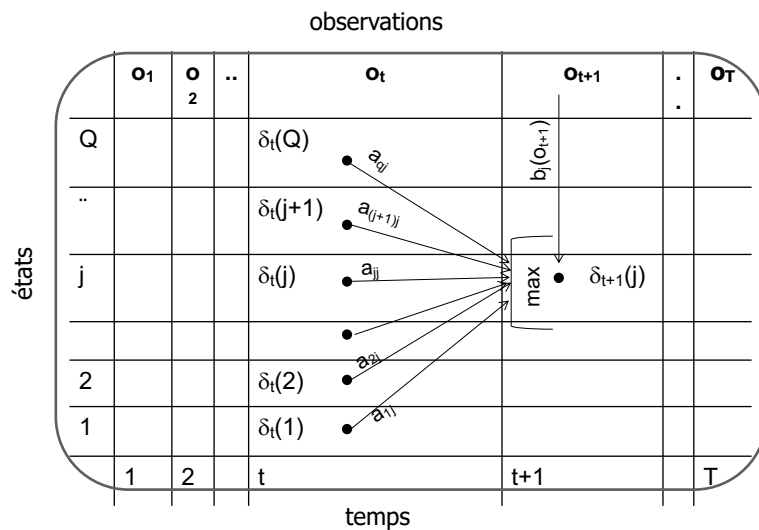
- backtrack

$$\hat{q}_t = \varphi(\hat{q}_{t+1}) \quad t = T-1, T-2, \dots, 1$$

Laurence Likforman-Telecom
ParisTech

31

31

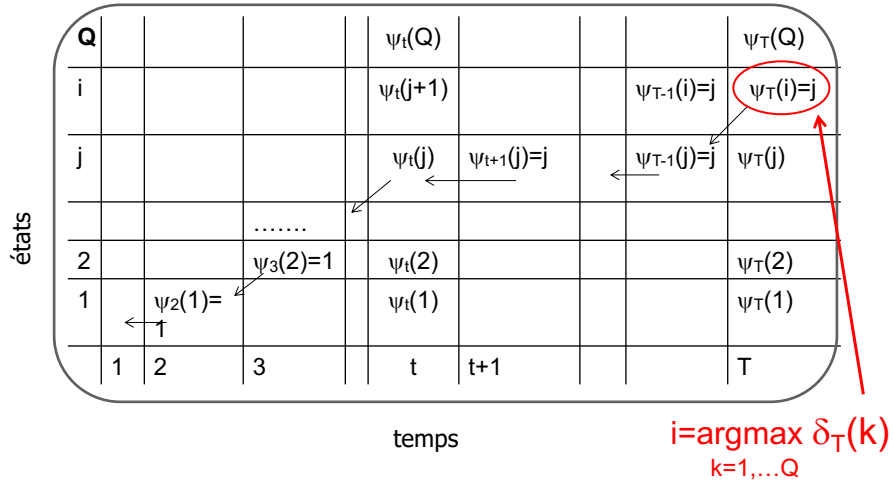


Laurence Likforman-Telecom ParisTech

32

32

backtrack



Laurence Likforman-Telecom
ParisTech

33

33

références

- M. Sigelle, Bases de la Reconnaissance des Formes: Chaînes de Markov et Modèles de Markov Cachés, chapitre 7, Polycopié Telecom ParisTech, 2012.
- L. Likforman-Sulem, E. Barney Smith, Reconnaissance des Formes: théorie et pratique sous matlab, Ellipses, TechnoSup, 2013.
- L. Rabiner, A tutorial on Hidden Markov Models and selected applications in Speech Recognition, proc. of the IEEE, 1989.

Laurence Likforman-Telecom
ParisTech

34

34