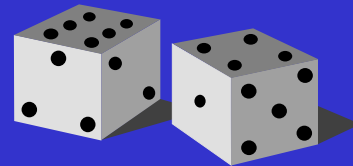


Modèles de Markov Cachés

*Modèles de Markov Cachés (Hidden
Markov Models (HMMs))*

Tarik AL ANI, Laboratoire
A2SI-ESIEE-Paris / LIRIS
e-mail : t.alani@esiee.fr



Tarik AL ANI, A2SI-ESIEE – Paris/LIRIS

Introduction

Thème : *Apprentissage artificiel* (Machine Learning): logiciel intelligent pour l'analyse des données.

Quoi : Trouver des modèles pour représenter les données.

Pourquoi : Pour effectuer une prédiction, pour comprendre, ...

Comment : Rechercher un «bon» modèle.

Modèles de Markov Cachés

- ***Apprentissage*** = Données & Représentation
+ Connaissance a priori
+ Procédure ou règle d'apprentissage
+ Exploration/Optimisation

Domaine Statistiques : Estimation des paramètres

Systemes et Contrôle : Identification des paramètres d'un système.

Modèles de Markov Cachés

- *HMMs* : *modèles statistiques* utiles pour des *données séquentielles*.
- Domaines d'application:
Intelligence Artificielle, Reconnaissance de formes, Reconnaissance de la parole, Identification et contrôle, Traitement du signal, Diagnostic, Analyse des séquences biologiques, Robotique, Finance, Analyses géopolitiques, ...

Modèles de Markov Cachés

- **Plusieurs variantes de HMMs**
 - ***HMMs conventionnels [Rabi 89]***
 - HMM Hybrides (HMMs+ANNs)
 - Input-Output HMMs (IOHMMs)
 - Variable-Length Markov Models (VLMMs)
 - Markov Switching Models (MSMs)
 - HMMs hiérarchiques (HHMMs)
 - HMMs Factoriels (FHMMs)
 - Weighted Transducers (WTs)
 - ...

Modèles de Markov Cachés

- Soit $\mathbf{y}_1^T = \{y_1, y_2, \dots, y_T\}$ une séquence d'observations de $t=1$ à $t=T$.
- Cette donnée séquentielle (séquence d'observations) en considération obéit à certaines propriétés.

Distribution de probabilité d'une séquence d'observations \mathbf{y}_1^T peut toujours être déterminée par

$$P(\mathbf{y}_1^T) = P(y_1) \prod_{t=2}^T P(y_t | y_1^{t-1}).$$



Modéliser une séquence d'observations dans laquelle la *distribution conditionnelle* de probabilité $P(y_t \setminus y_1^{t-1})$ d'une variable y_t scalaire ou vecteur de dimension K observée au temps t dépend de l'historique des valeurs précédentes y_1^{t-1} est en général un problème pas facile en pratique.

Modèles de Markov

- Un modèle de Markov d'ordre k est une distribution de probabilité sur une séquence de variables

$$q_1^t = \{q_1, q_2, \dots, q_t\}$$

avec la propriété d'indépendance conditionnelle suivante:

$$P(q_t \setminus q_1^{t-1}) = P(q_t \setminus q_{t-k}^{t-1}).$$

Puisque q_{t-k}^{t-1} résume toutes les informations relevées du passé, q_t est appelée en général variable d'état.

Modèles de Markov Cachés

Grâce à la propriété précédente, la distribution conjointe de la séquence entière peut être décomposée en produits

$$P(q_1^T) = P(q_1^k) \prod_{t=k+1}^T P(q_t \setminus q_{t-k}).$$

Cas spécial : **Modèle de Markov d'ordre 1**

$$P(q_1^T) = P(q_1) \prod_{t=2}^T P(q_t \setminus q_{t-1}).$$

Ce modèle est complètement spécifié par l'**état initial** $P(q_1)$ et par les **probabilités de transition**

$$P(q_t \setminus q_{t-1}).$$

Modèles de Markov Cachés

Probabilités de transition

$$a_{q_t q_{t-k}} = P(q_t \setminus q_{t-k}^{t-1})$$

– *Homogène* : La même $\forall t$.

e.g. Modèles de Markov d'ordre 1:

$$P(q_t=j \setminus q_{t-1}=i) = P(q_t=j \setminus q_{t-1}=i)$$

- Les séquences possèdent des transitions temporelles invariantes
- Un nombre réduit de paramètres
- Le modèle peut être entraîné sur des séquences de certaines longueurs et généralisé à des séquences de longueurs différentes.

Modèles de Markov Cachés

– *Non-homogène* : dépendent de t .

e.g. Modèles de Markov d'ordre 1:



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

Les séquences possèdent des transitions temporelles variantes.



Les probabilités de transition ne sont pas les paramètres directs du modèle mais sont plutôt obtenus comme une fonction « paramétrée » de l'état précédent et d'autres paramètres de conditionnement.

Modèles de Markov Cachés

-  Les mêmes avantages des probabilités de transition homogènes.
-  Plus de capacité d'adaptation à certains changements de la dynamique observée dans les différentes parties de la séquence.

Nature de l'état

- *état discret* : distribution discrète (beaucoup d'applications)
- *état continu* : distribution continue (systèmes à espace d'état continu).

Dans la suite de cet exposé nous considérons uniquement le cas de l'état discret.

Modèles de Markov Cachés

Ordre de la chaîne de Markov

Les modèles Markoviens d'ordre k sont non pratiques.

$$q_{t-k}^{t-1}$$

Exemple : pour la variable d'état $q_t \in \{1, 2, \dots, N\}$, le nombre nécessaire de paramètres pour représenter les probabilités de transition est $O(N^{k+1})$.

Restriction : Utilisation d'une petite valeur de k . La majorité des séquences observées en pratique ne satisfait pas l'hypothèse de Markov d'ordre k faible !!



Modèles de Markov Cachés

Nous ne supposons pas que la séquence observée possède une propriété Markovienne (d'ordre k faible),

Mais nous supposons qu'une autre variable non observée (*variable d'état*) qui est liée à cette séquence et qui possède la propriété de Markov d'ordre 1.



Modèles de Markov Cachés

Les HMMs cette propriété : la séquence du passé peut être résumée par une variable aléatoire *non observée* (*cachée*) appelée *variable d'état*, qui transmet toute l'information y_1^{t-1} nécessaire pour décrire la distribution de l'observation future y_t .

Ceci est précisément l'idée de départ de HMMs.

Modèles de Markov Cachés

- *Le problème d'apprentissage de Modèles Markoviens* :

Etant donné

- un *ensemble de données d'apprentissage* $O = \{O_1, O_2, \dots, O_W\}$

de L séquences d'observation $O_w = \{y_1, y_2, \dots, y_T\}$

y_1 est un scalaire ou un vecteur de dimension K .

- un *critère* ℓ pour définir la qualité du modèle sur un ensemble de données (lier O et un modèle λ_m à un scalaire à valeur réelle), choisir $\lambda_m, m \in \{1, 2, \dots, M\}$ à partir d'un ensemble de modèles $A = \{\lambda_m\}$, pour maximiser (ou minimiser) la valeur espérée de ce critère sur une nouvelle donnée d

Modèles de Markov Cachés

- Dans certaines applications il y a seulement une séquence d'observations $O_1 = O_t = y^t = \{y_1, y_2, \dots, y_t\}$ et la nouvelle donnée est, simplement une continuité des données d'apprentissage, (e.g., prédiction des séries temporelles, « économétrie »).
- Dans d'autres applications il y a un grand nombre de séquences d'apprentissage de longueurs différentes, e.g., base de données en traitement de la parole : des milliers ou des dizaines de milliers de séquences).

Modèles de Markov Cachés

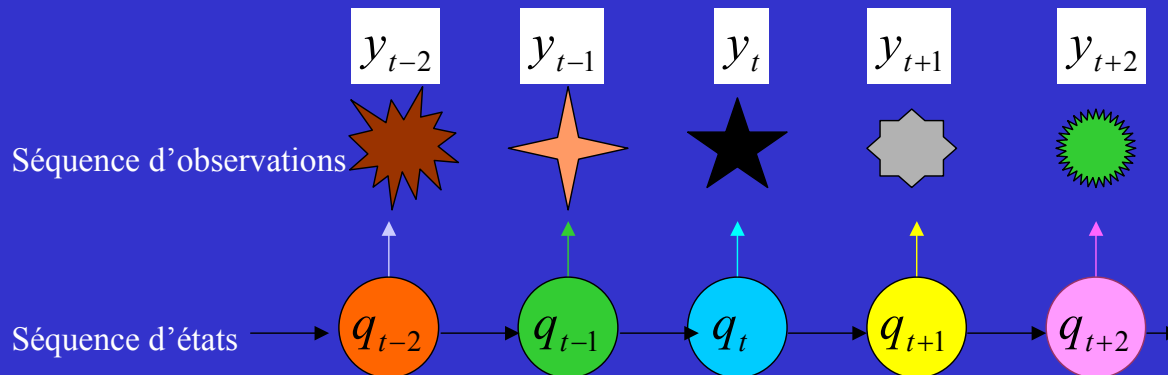
Exemple d'un modèle de Markov d'ordre 1:

$$P(y_t \setminus q_1^t, y_1^{t-1}) = P(y_t \setminus q_t) \quad (1)$$

$$P(q_{t+1} \setminus q_1^t, y_1^t) = P(q_{t+1} \setminus q_t) \quad (2)$$

(1) est la prédiction de y_t basée sur q_t uniquement.

(2) est la prédiction de q_{t+1} basée sur q_t uniquement.



Modèles de Markov Cachés

La **distribution conjointe** de la variable d'état caché et de la variable observée est alors

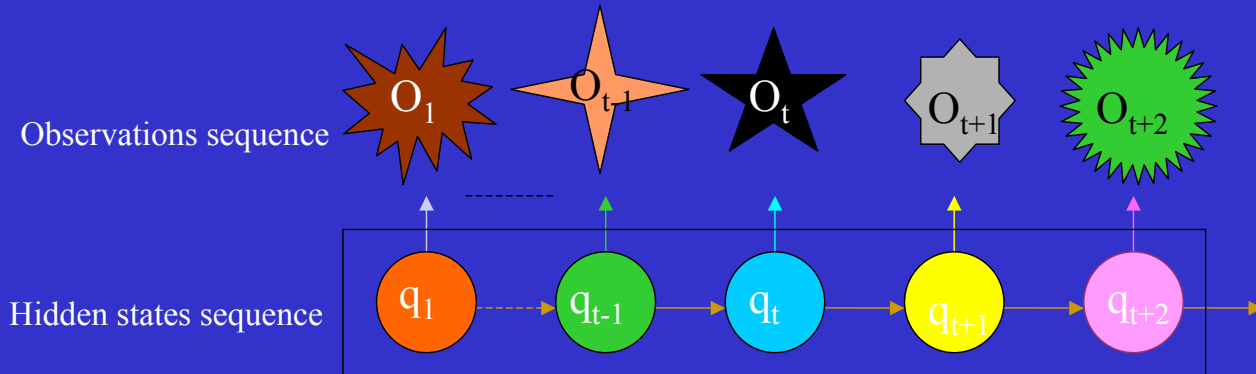
$$P(y_1^T, q_1^T) = P(q_1) \prod_{t=1}^{T-1} P(q_{t+1} \setminus q_t) \prod_{t=1}^T P(y_t \setminus q_t). \quad (3)$$

Cette distribution est complètement spécifiée en termes de:

1. **probabilités de l'état initial** $P(q_1)$,
2. **probabilités de transition** $P(q_{t+1} \setminus q_t)$ et,
3. **probabilités d'émission** $P(y_t \setminus q_t)$.

Modèles de Markov Cachés

Modèles de Markov Cachés (Hidden Markov Models (HMMs))



Processus dynamique, N states ($q_t \in S = \{1, 2, \dots, N\}$)

$$\Pi = [\pi_1, \pi_2, \dots, \pi_N];$$

$$\pi_i = p(q_1 = i), i \in S$$

$$a_{ij} = p(q_{t+1} = j \mid q_t = i), A = [a_{ij}], i, j \in S$$

$$b_i(y_t) = P(y_t \mid q_t = i) = N(M_i, \Sigma_i),$$

Connaissance a priori de la structure du modèle

- **Etat initial** : Dans beaucoup d'applications où la variable d'état est discrète, toutes les séquences d'état sont supposées de commencer par un état initial j

$$q_t \in \{1, 2, \dots, N\}, \quad \pi_i = P(q_1 = i) = 1, \quad i = j, \\ = 0, \quad i \neq j, \quad \sum_i \pi_i = 1,$$

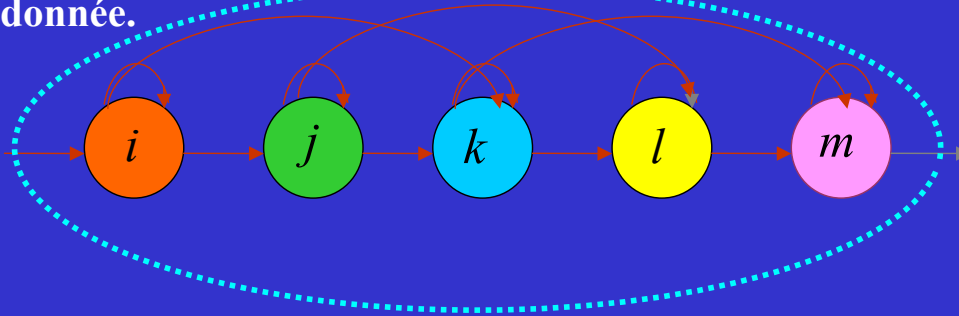
et terminer dans un état final.

Modèles de Markov Cachés

- **Topologie** : Certaines probabilités de transitions sont forcées à zéros. Un nœud représente la valeur de variable d'état q_t . Un arc représente une transition entre deux variables d'état avec une probabilité non nulle.

Modèles de Markov Cachés

- Exemple d'une topologie gauche-à-droite d'un HMM.
 - Reconnaissance de la parole : pour représenter la distribution des séquences acoustiques associées à une unité de parole (e.g. phonème, mot).
 - évolution d'un système dynamique : pour représenter la distribution des séquences d'états associées à une trajectoire donnée.



Modèles de Markov Cachés

- Dans une variante liée au modèle précédent, les émissions des observations ne dépendent pas uniquement de l'état courant mais aussi de l'état précédent (i.e. transitions) :

$$P(y_1^T, q_1^T) = P(q_1) P(y_1 | q_1) \prod_{t=2}^T P(q_t | q_{t-1}) P(y_t | q_t, q_{t-1}). (3)'$$

Modèles de Markov Cachés

Nous sommes intéressés à la distribution $P(y_1^T)$.



Problème: q_1^T n'est pas observé !



Marginaliser la distribution conjointe:

$$P(y_1^T) = \sum_{q_1^T} P(y_1^T, q_1^T)$$



Un nombre exponentiel de termes.



Le calcul vers l'avant (Forward)

Calcul récursif : basé sur la factorisation des probabilités qui prennent avantage de la propriété de Markov d'ordre 1 (equ. (1) et (2)).

La récurions n'est pas effectuée sur $P(y_1^t)$ elle même mais sur la probabilité $P(y_1^t, q_1^t)$ d'observer une certaine séquence pendant que l'état prend une valeur particulière à la fin de cette séquence.

Modèles de Markov Cachés

$$\begin{aligned} P(y_1^t, q_t) &= P(y_t, y_1^{t-1}, q_t) P(y_1^{t-1}, q_t) & (4) \\ &= P(y_t \setminus q_t) \sum_{q_{t-1}} P(y_1^{t-1}, q_t, q_{t-1}) \\ &= P(y_t \setminus q_t) \sum_{q_{t-1}} P(q_t \setminus q_{t-1}, y_1^{t-1}) P(y_1^{t-1}, q_{t-1}) \end{aligned}$$

Nous obtenons alors

$$P(y_1^t, q_t) = P(y_t \setminus q_t) \sum_{q_{t-1}} P(q_t \setminus q_{t-1}) P(y_1^{t-1}, q_{t-1}) \quad (5)$$

Modèles de Markov Cachés

- La récursions peut être initialisée avec $P(y_1, q_1) = P(q_1)P(y_1 \setminus q_1)$ en utilisant les probabilités de l'état initial $P(q_1)$.
- elle peut être utilisée pour un modèle homogène ou non homogène.
- les probabilités peuvent être conditionnées par d'autres variables.
- elle est centrale à plusieurs algorithmes pour les HMMs.
- Elle permet de calculer la *fonction de vraisemblance (l)*:

$$l(\lambda) = \prod_w P(y_1^{T_w}(w) \setminus \lambda),$$

où λ regroupe les paramètres du modèle qui peuvent être ajustés pour maximiser l sur toutes les séquences d'apprentissage $y_1^{T_w}(w)$.

Modèles de Markov Cachés

– Le coût du calcul est $O(Tm)$

où

T est la longueur de la séquence

m est le nombre de probabilités de transition non nulle à chaque t,

Par les connaissances a priori de la structure (topologie) du modèle, $m \ll N^2$ dans beaucoup de cas en pratique. N est le nombre d'états.

Modèles de Markov Cachés

Une fois $P(y_1^T, q_T \setminus \lambda)$ calculée, on peut immédiatement calculer la fonction de vraisemblance pour chaque séquence : $P(y_1^T \setminus \lambda)$.

$$P(y_1^T \setminus \lambda) = \sum_{q_T} P(y_1^T, q_T \setminus \lambda)$$

Remarque : Parfois, pour alléger l'écriture, nous supprimons le conditionnement des probabilités par le modèle λ .

Modèles de Markov Cachés

Choix des distributions

- Probabilités de transition

- Cas d'un état discret :

- Matrice des probabilités de transition A d'un modèle homogène :

$$A = [a_{ij}] = [P(q_{t+1} = i \mid q_t = j)], a_{ij} \geq 0, \sum_i a_{ij} = 1, \forall j.$$

- Cas d'un état continu (modèles à espace d'état continu)

Dans ce cas la distribution de l'état futur est, d'habitude, une fonction Gaussienne dont la moyenne est une fonction de l'état précédent.

- **Probabilités de l'émission (observation)** y_t
 - Cas d'une **observation discrète** (DHMMs):
 - **Matrice des probabilités d'observations** **B** d'un modèle homogène

$$B = [b_j(v_k)] = [P(y_t = v_k \mid q_t = j)], \quad \sum_k b_j(v_k), \forall j$$

où

$v_k \in V = \{v_1, v_2, \dots, v_M\}$ l'ensemble des symboles discrets

Modèles de Markov Cachés

- Cas d'une observations vectorielles continues,

Il existe plusieurs techniques, notamment:

1. Quantification vectorielle

Assigner chaque vecteur y_t à une valeur discrète $quantize(y_t)$ et utiliser $P(quantize(y_t) \setminus q_t)$ comme la probabilité d'émission.

2. Observation continue (CHMMs):

- Distribution Gaussienne vectorielle

$$b_j(y_t) = P(y_t \mid q_t = j) = N(y_t; \mu_j, \Sigma_j), \int_{-\infty}^{+\infty} b_j(y_t) dy_t = 1, \forall j$$

où

$N(y_t; \mu_j, \Sigma_j)$ est la probabilité d'observer le vecteur y_t sous la distribution Gaussienne paramétrée par le vecteur des valeurs moyennes conditionnées par l'état j et par la de covariance conditionnée aussi par le même état.

- Mélange de distributions Gaussiennes

$$b_j(y_t) = P(y_t \mid q_t = j) = \sum_m c_{mj} N(y_t; \mu_j, \Sigma_j), \text{ où } c_{mj}, \sum_m c_{mj} = 1.$$

3. Distributions pseudo-continues (semi-continuous) (*SCHMMs*)

Dans ce cas, les paramètres spécifiques à chaque état sont uniquement les poids du mélange:

$$b_j(y_t) = P(y_t \mid q_t = j) = \sum_m c_{mj} N(y_t; \mu_m, \Sigma_m), \text{ où } c_{mj}, \sum_m c_{mj} = 1.$$

où c_{mj} jouent un rôle qui est similaire aux $b_j(v_k)$ dans le cas discret.

Estimation des paramètres du modèle
(*apprentissage*)

Le *modèle initial*

$\lambda_0 = (\Pi, A, B)$ dans le cas discret

$\lambda_0 = (\Pi, A, \mu, \Sigma)$ dans le cas continu

où $\Pi = [\pi_1, \pi_2, \dots, \pi_N]$ est le vecteur des probabilités de l'état initial ($\pi_i = P(q_t = i)$)

A est la matrice de transition

B est la matrice d'émissions

$\mu = [\mu_i]$ est le vecteur des valeurs moyennes conditionnées par l'état

$\Sigma = [\Sigma_i]$ est la matrice composée des matrices de covariances conditionnées par l'état.

Modèles de Markov Cachés

Le problème d'estimation des paramètres

Etant donné un modèle initial λ_0 et un ensemble de données d'apprentissage, comment peut-on ajuster les paramètres de ce modèle?

Ensemble fini d'apprentissage : Plusieurs méthodes d'optimisations de nature itératives pour obtenir un maximum local de $P(y_1^t \mid \lambda)$ sur l'ensemble de séquences d'apprentissage O . Le modèle est ajusté à chaque itération jusqu'à ce qu'un critère de fin soit satisfait.

Modèles de Markov Cachés

Trois étapes :

1. Apprentissage d'un modèle λ_k de chaque événement k (e. g. respiration normale, ronflement) en utilisant plusieurs séquences de données d'apprentissage d'horizon T : $y_1^T = y_1, y_2, \dots, y_T$ appartenant à cet événement.
2. Segmentation d'une séquence de données en une séquence optimale d'états liés à un événement $Q^* = q_1 q_2 q_3, \dots, q_T$. Puis l'interprétation de ces états peut être fait par un expert (e. g. : q_1 inspiration, q_2 : expiration).
3. Détection Hors-ligne/En-ligne des états pathophysiologiques en utilisant tous les modèles optimisés $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_K$.

Modèles de Markov Cachés

1. Apprentissage des HMMs (Estimation de paramètres du Model)

Modèle de l'événement
k à l'itération it

N états ($q_t = S \in \{1, 2, \dots, N\}$)

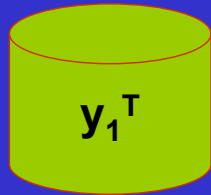
$\Pi = [\pi_1, \pi_2, \dots, \pi_N]$; $\pi_i = p(q_1 = i)$

$a_{ij} = p(q_{t+1} = j \mid q_t = i)$, $A = [a_{ij}]$

$i, j \in S$

$b_i(y_t) = P(y_t \mid q_t = i) = N(M_i, \Sigma_i)$,

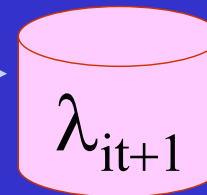
$\lambda_k^* = \{\Pi, A, M, \Sigma\}$



Séquences d'observations
pour l'apprentissage

y_1, y_2, \dots, y_T
de l'événement k.

27/01/2006



Modèle Optimisé de
l'événement k (λ_k^*)

Modèles de Markov Cachés

Une liste non exhaustive des algorithmes d'apprentissage:

- *Algorithme de Baum-Walsh* (utilisée couramment à cause de sa propriété de convergence relativement rapide)

Cette algorithme est basé sur un autre algorithme appelé EM (Estimation-Maximization).

Il existe plusieurs variants de cette algorithme mais il sont tous basés sur le principe de maximum de vraisemblance, e.g. :

- *Maximum Mutual Information (MMI)*
- *Minimum Discrimination Information (MDI)*

Principe

Un nouveau ensemble de paramètres $\tilde{\lambda}$ est choisi tel que $P(y_1^T \setminus \lambda)$ est maximisée pour la séquence donnée y_1^T .

Maximiser une fonction auxiliaire

$$Q(\lambda_t, \tilde{\lambda}) = \sum_{q_1^T} P(q_1^T \setminus y_1^T, \lambda) \log P(y_1^T, q_1^T \setminus \tilde{\lambda}) \text{ sur } \tilde{\lambda}$$



Maximiser la vraisemblance

$$P(y_1^T \setminus \tilde{\lambda}) \geq P(y_1^T \setminus \lambda)$$



Un point critique

Modèles de Markov Cachés

Propriétés



Converge relativement rapidement



Peut être utilisé en ou hors ligne



Le modèle obtenu est sensible à l'estimation initiale.

Des paramètres initiaux à valeurs nulles restent les mêmes à la fin de l'apprentissage.

– **Algorithme de Viterbi**

Cette algorithme est basé sur le critère maximum a posteriori.

Principe Partant d'un état initial $\tilde{\lambda}_r$ calculer

$$\tilde{\lambda}_{r+1} = \arg \max_{\tilde{\lambda}} \max_{q_1^T} P(y_1^T, q_1^T \mid \tilde{\lambda}_r)$$

Propriété



Converge rapidement,



Peut être utilisé hors ou en ligne.

– *Algorithme de Recuit simulé* [Hama 96]

Principe

Etant donné une séquence d'observations y_1^T ,

1. une séquence d'état est générée au hasard,
2. une observation est assignée à chaque état pour calculer la moyenne et la variance,
3. en utilisant ses valeurs, la probabilité (*coût initial*) de générer y_1^T avec la séquence sélectionnée est calculée,

Modèles de Markov Cachés

4. un état dans la séquence est modifié aléatoirement , retour à l'étape 3 pour calculer la probabilité avec la nouvelle séquence d'état générée après avoir changé la moyenne et la variance des états modifiés.

Si la nouvelle probabilité est plus grande que la précédente, ou si la **probabilité d'acceptation**

$$P_{accept} = \exp((P_{new} - P_{old}) / Temp)$$

est assez grande, alors la modification est sauvegardée.

Temp est appelé **Température**.

Pour des valeurs faibles de Temp, uniquement les meilleures solutions sont acceptées.

Modèles de Markov Cachés

Propriétés



La température finale $Temp_f$ est calculée automatiquement à partir des considérations théoriques



Converge lentement (mais sûrement !!)



Peut être utilisé uniquement hors ligne

Modèles de Markov Cachés

- Estimation de la séquence d'état (*Algorithme de Viterbi*)

Il est utile d'estimer la séquence optimale d'états qui correspond à une séquence donnée d'observations.

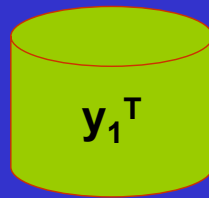
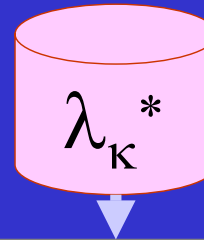
Ceci peut être réalisé par l'algorithme de Viterbi basé sur la programmation dynamique stochastique (récursif):

$$q_1^{T*} = \arg \max_{q_1^T} P(q_1^T \setminus y_1^T, \lambda) = \arg \max_{q_1^T} P(q_1^T, y_1^T \setminus \lambda)$$

Modèles de Markov Cachés Estimation de la séquence d'états

$$Q^* = \arg \max_Q [p(Q, y_1^T | \lambda_k^*)]$$

Modèle de l'événement k



Séquences d'observations
pour l'apprentissage
 y_1, y_2, \dots, y_T
de l'événement k.

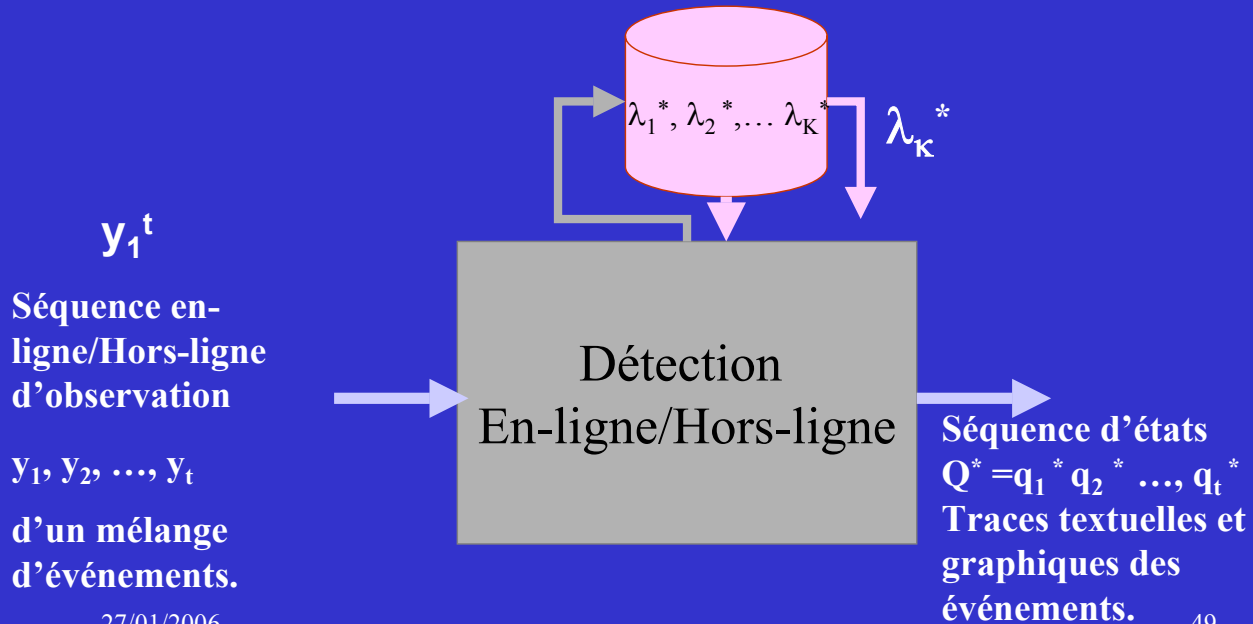
**Programmation Dynamique
Algorithme de Viterbi**

Séquence d'états
 $Q^* = q_1 q_2 q_3, \dots, q_T$
de l'événement k.

Modèles de Markov Cachés
Détection En-ligne/Hors-ligne d'un événement.

$$\lambda^* = \arg \max_k [p(y_1^t | \lambda_k)] \quad q_t^* = \arg \max_i [p(y_1^t, q_t=i | \lambda^*)]$$

Modèles des événements respiratoires



Modèles de Markov Cachés

Conclusions

- Les HMMs sont des outils mathématiques puissants et très élégants qui peuvent être appliqués dans beaucoup de domaines.
- Leurs récentes extensions notamment au temps réel peuvent les rendre encore importants pour beaucoup de tâches d'apprentissage.
- Plusieurs problèmes d'ordres théoriques et pratiques restent à explorer, ce qui donne des challenges pour la recherche future.
- Liste de mes publications en HMM :
<http://www.esiee.fr/~alanit/>
- Hidden Markov Models scilab toolbox :
<http://scilabsoft.inria.fr/contribution/displayCategory.php?category=MODELING%20AND%20CONTROL%20TOOLS>