Méthodes probabilistes pour le traitement automatique de la langue

LI 031

Table des matières

Chapitre 1

Modèles de Markov cachés

1.1 Chaînes de Markov

1.1.1 Processus stochastique

Un **processus stochastique** (ou processus aléatoire) est une séquence $X_1, X_2 \dots X_N$ de variables aléatoires fondées sur le même ensemble fondamental S. Les valeurs possibles des variables aléatoires sont appelées les états possibles du processus. La variable X_t représente l'état du processus au temps t (on dit aussi l'observation au temps t). Les différentes variables aléatoires ne sont en général pas indépendantes les unes des autres. Ce qui fait réellement l'interêt des processus stochastiques est la dépendance entre les variables aléatoires.

Pour spécifier entièrement un processus stochastique, il suffit de spécifier :

1. la loi de probabilité de la première variable alétaoire X_1, qui spécifie donc l'état du processus lors de la première observation.

2. pour toute valeur de $t > 1$ la probabilité conditionnelle $P(X_t = j | X_1 = i_1, \dots, X_{t-1} = i_{t-1})$

1.1.2 Propriété de Markov

Une **chaîne de Markov** est un type particulier de processus stochastique qui vérifie deux conditions :

– L'état au temps t du processus ne dépend que de son état au temps $t-1$:

$$P(X_t = j | X_1 = i_1, \dots, X_{t-1} = i_{t-1}) = P(X_t = j | X_{t-1} = i_{t-1})$$

– La probabilité de passage d'un état i à un état j ne varie pas avec le temps :

$$\forall t, 1 < t \leq N, P(X_t = j | X_{t-1} = i) = C$$

Un processus de Markov peut être décrit par une **matrice de transition** T telle que :

$$T(i,j) = P(X_t = j | X_{t-1} = i), 1 < t \leq N$$
$$\text{avec } T(i,j) \geq 0, \ \forall i, j$$
$$\text{et } \sum_{j=1}^{N} T(i,j) = 1 \ \forall i$$

Il faut définir de plus l'état du processus à l'instant 1 donc la loi de probabilité, notée π de la variable X_1 :

$$\pi(i) = P(X_1 = i)$$

On peut éviter le recours à celle loi en imposant que le processus débute toujours dans le même état s_0, par exemple et en utilisant les transitions depuis cet état pour représenter les probabilités π.

Un processus de Markov peut aussi être représenté par un automate fini où chaque état du processus est représenté par un état de l'automate. Les transitions sont représentées par des flèches reliant les états, ces flèches sont étiquetées par les probabilités correspondantes.

Exemple :
On admet que le fait qu'il ait plu ou non un jour donné est la seule considération à prendre en compte pour prévoir s'il pleuvra le lendemain. Plus précisément, s'il pleut aujourd'hui, il pleuvra demain aussi avec une probabilité de α et s'il ne pleut pas aujourd'hui la probabilité qu'il pleuve demain est β.

On convient de dire que le système est dans l'état 1 s'il pleut et 2 s'il ne pleut pas. La situation peut être représentée par une chaîne de Markov à deux états dont la matrice de transition est :

$$\begin{vmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{vmatrix}$$

De plus, la probabilité que le processus soit dans l'état 1 à l'instant 1 est égale à γ.

Le même processus peut être représenté par l'automate de la figure 1.1.

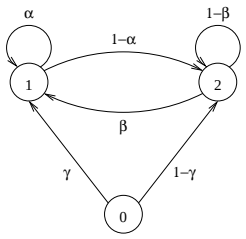

FIG. 1.1 – Représentation d'un processus de Markov sous la forme d'un automate fini

1.1.3 Probabilité d'une suite d'observations

Les propriétés de Markov permettent de calculer simplement la probabilité qu'une suite d'états particulière de longueur T soit observée (la loi de probabilité conjointe de $(X_1, X_2, \ldots X_T)$) :

$$\begin{aligned} p(X_1, X_2, \ldots, X_T) &= p(X_1)p(X_2|X_1)p(X_3|X_1, X_2)\ldots p(X_T|X_1, \ldots, X_{T-1}) \text{(règle de multiplication)} \\ &= p(X_1)p(X_2|X_1)p(X_3|X_2)\ldots p(X_T|X_{T-1}) \text{(hypothèse de Markov)} \end{aligned}$$

Exemple :
Etant donné le processus de Markov de l'exemple précédent, la probabilité d'avoir trois jours consécutifs de pluie est égale à :

$$\begin{aligned} p(X_1 = 1, X_2 = 1, X_3 = 1) &= p(X_1 = 1)p(X_2 = 1|X_1 = 1)p(X_3 = 1|X_2 = 1) \\ &= \gamma \times \alpha \times \alpha \\ &= \gamma\alpha^2 \end{aligned}$$

Notations

Afin d'alléger les notations, on omettra de spécifier les variables aléatoires dans l'écriture de la probabilité d'une suite d'états. On notera, par exemple, $p(1, 1, 1)$ la probabilité $p(X_1 = 1, X_2 = 1, X_3 = 1)$.

1.2 Modèles de Markov Cachés

Dans les chaînes de Markov telles que présentées ci-dessus, les observations correspondent aux états du processus. Dans un modèle de Markov caché, on ne peux observer directement les états du processus, mais des symboles (appelés aussi *observables*) émis par les états selon une certaine loi de probabilité. Au vu d'une séquence d'observation on ne peux par conséquent savoir par quelle séquence d'états (ou *chemin*) le processus est passé, d'où le nom de modèles de Markov cachés (MMC). On distinguera donc dans cette section le processus $X = X_1, X_2, \ldots, X_T$ qui représente comme dans la section précédente l'évolution des états du MMC et le processus $O = O_1, O_2, \ldots, O_T$ qui représente la suite des symboles émis par le MMC.

1.2.1 Eléments d'un MMC

Un MMC est défini par un quintuplet $\langle S, A, \pi, T, E \rangle$ où :

- S est l'ensemble des états : $\{1, \ldots, N\}$
- A est l'alphabet des symboles émis par les états : $\{a_1, \ldots, a_M\}$
- π est la loi de probabilité de l'état initial $\pi(i) = P(X_1 = i)$. π étant une loi de probabilité, on a :

$$\sum_{i=1}^{N} \pi(i) = 1$$

- T est la matrice des probabilités de transition d'un état vers un autre. La probabilité de transition d'un état i vers un état j ($P(X_t = j | X_{t-1} = i)$) est notée $T(i, j)$. Comme dans le cas des chaînes de Markov, la somme des probabilités des transitions émanant d'un état vaut 1 :

$$\sum_{j=1}^{N} T(i, j) = 1, \ \forall \ i \in S$$

- E est la matrice des probabilités d'émission des symboles de A pour chaque état. La probabilité que l'état i émette le symbole j ($P(O_t = j | X_t = i)$) est notée $E(i, j)$. Les probabilités d'émission de symboles de A pour chaque état du MMC constituent une loi de probabilité, on a par conséquent :

$$\sum_{j=1}^{M} E(i, o_j) = 1, \ \forall \ i \in S$$

L'ensemble constitué des probabilités initiales, des probabilités de transition et d'émission d'un MMC λ est souvent appelé les *paramètres* λ.

Exemple :

Le MMC $\lambda_1 = \langle \{1, 2, 3\}, \{a, b, c\}, \pi, T, E \rangle$ avec :

$$\begin{array}{lll} E(1,a) = 0,6 & E(2,a) = 0 & E(3,a) = 0,3 \\ E(1,b) = 0,2 & E(2,b) = 0,5 & E(3,b) = 0 \\ E(1,c) = 0,2 & E(2,c) = 0,5 & E(3,c) = 0,7 \end{array} \quad \text{et} \quad \begin{array}{lll} T(1,1) = 0,3 & T(2,1) = 0,6 & T(3,1) = 0,2 \\ T(1,2) = 0,2 & T(2,2) = 0,1 & T(3,2) = 0,4 \\ T(1,3) = 0,5 & T(2,3) = 0,3 & T(3,3) = 0,4 \end{array}$$

et

$$\pi(1) = 1 \quad \pi(2) = 0 \quad \pi(3) = 0$$

est représenté graphiquement dans la figure 1.2

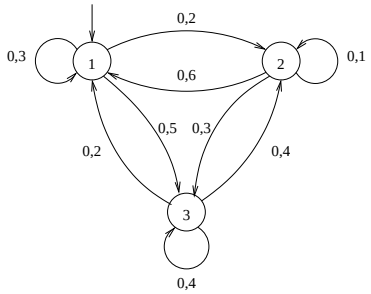

FIG. 1.2 – Représentation graphique du MMC λ

1.2.2 Calcul de la probabilité d'une séquence d'observations

Etant donné un MMC $\lambda = \langle S, A, \pi, T, E \rangle$, la suite d'observation $o = o_1 o_2, \ldots, o_T$ peut généralement être générée en suivant différents chemins dans le MMC [1]. La probabilité que λ émette la séquence o est égale à la somme des probabilités que la séquence o soit émise en empruntant les différents chemins pouvant émettre o. Ce raisonnement correspond en fait à l'application de la formule des probabilités totales à la probabilité $P(o)$ [2] :

$$P(o) = \sum_{x \in \mathcal{C}_T} P(o|x)P(x) \tag{1.1}$$

où $\mathcal{C}_T$ est l'ensemble des séquences de T états de λ et $x = x_1, \ldots, x_T$ ($x_i \in S,\ 1 \leq i \leq T$) une de ces séquences [3].

la probabilité conditionnelle que o soit générée lorsque λ passe successivement par la séquence d'états $x = x_1, \ldots, x_T$ est le produit des probabilités que l'état atteint à l'instant t (x_t) émette le symbole observé à cet instant (o_t) :

$$P(o|x) = \prod_{t=1}^{T} E(x_t, o_t)$$

et la probabilité que le MMC suive une séquence particulière d'états x est le produit des probabilités que λ passe de l'état x_t à l'état x_{t+1} entre les instants t et $t+1$, comme dans un modèle de Markov *visible* :

$$P(x) = \pi(x_1) \prod_{t=1}^{T-1} T(x_t, x_{t+1})$$

En remplaçant $P(o|x)$ et $P(x)$ dans l'équation 1.1, il vient :

[1] Si l'on voit le MMC comme un automate à états finis, ceci revient à dire que l'automate est non déterministe.

[2] De la même façon que nous avons noté dans la section précédente $p(x_1, x_2, x_3)$ la probabilité qu'une chaîne de Markov passe successivement par les états x_1, x_2 et x_3, on notera $p(o_1, o_2, o_3)$ la probabilité qu'un modèle de Markov cachée émette successivement les trois symboles o_1, o_2 et o_3. La probabilité conditionnelle $P(o|x)$ correspond donc à $P(O_1 = o_1, O_2 = o_2, \ldots O_T = o_T | X_1 = x_1, X_2 = x_2 \ldots X_T = x_t)$.

[3] Avant d'appliquer la formule des probabilités totales, il faut s'assurer que $\sum_{x \in \mathcal{C}_T} P(x) = 1$, ce qui qui se démontre facilement.

$$P(o) = \sum_{x \in \mathcal{C}_T} \pi(x_1) \times \prod_{t=1}^{T-1} E(o_t, x_t) T(x_t, x_{t+1}) \times E(o_T, x_T)$$

Malheureusement, le calcul ci-dessus est particulièrement inefficace, il nécessite dans le cas général (où tous les états sont reliés entre eux par une transition et chaque état peut émettre chacun des N symboles) $2 \times T \times N^T$ multiplications (N^T chemins et $2T$ multiplications à effectuer par chemin.).

C'est la raison pour laquelle on a recours à une méthode programmation dynamique pour effectuer ce calcul. Cette méthode repose sur la représentation, sous forme d'un *treillis*, de l'évolution du MMC ayant donné lieu à une suite d'observables $o_1 \dots o_k$, comme l'illustre la figure 1.3.

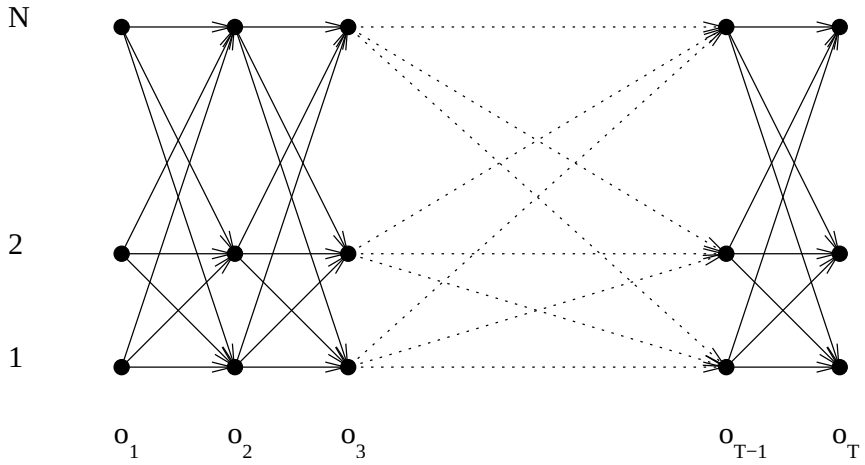

FIG. 1.3 – Représentation de l'évolution d'un MMC sous forme d'un treillis

On associe à chaque sommet (i, t) du treillis la variable $\alpha(i, t)$ qui correspond à la probabilité de se trouver dans l'état i du MMC λ à un instant t, ayant observé la suite $o_1 \dots o_{t-1}$:

$$\alpha(i, t) = P(o_1 \dots o_{t-1}, X_t = i)$$

L'avantage du treillis réside dans le fait qu'il est possible de représenter au niveau d'un sommet (i, t) de ce dernier des informations portant sur l'ensemble des chemins menant à l'état i à l'instant t tout en ayant observé la séquence $o_1 \dots o_{t-1}$. Dans notre cas, cette information est la somme des probabilités de ces chemins.

Cette particularité permet de calculer la probabilité de se trouver dans un état quelconque à un instant t en fonction de la probabilité de se trouver dans les différents états à l'instant $t - 1$, c'est l'étape récursive de l'algorithme décrit ci-dessous.

Algorithme de calcul de $P(o)$:

1. Initialisation :

$$\alpha(i,1) = \pi(i),\ 1 \leq i \leq N$$

2. Etape récursive :

$$\alpha(j,t+1) = \sum_{i=1}^{N} \alpha(i,t)E(i,o_t)T(i,j),\ 1 \leq t < T-1,\ 1 \leq j \leq N$$

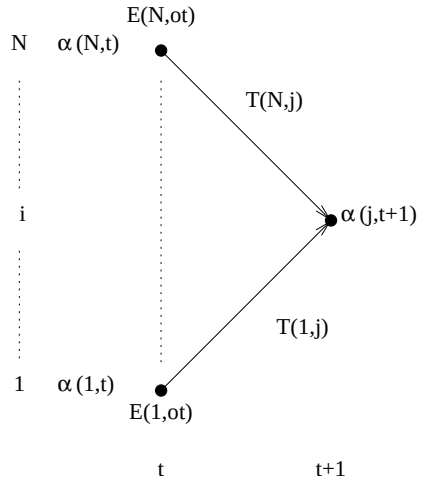

FIG. 1.4 – Calcul de $\alpha(j,t+1)$

3. Calcul de la probabilité totale :

$$P(o) = \sum_{i=1}^{N} \alpha(i,T)E(i,o_T)$$

Cette façon de calculer $P(o)$ est bien plus économique puisqu'elle n'exige (dans le cas général) que $2N^2T$ multiplications : $N \times T$ sommets et $2N$ multiplications par sommet.

La procédure de calcul de $P(o)$ présentée ci-dessus est appelée quelquefois procédure *forward* (en avant) car le calcul de la probabilité à un instant t est effectué à partir de la probabilité à un instant $t-1$, en parcourant le treillis de la gauche vers la droite. Il est aussi possible d'effectuer le calcul dans l'ordre inverse, où la probabilité à un instant t est calculée à partir de la probabilité à l'instant $t+1$. Ce mode de calcul, moins intuitif, est aussi appelé *backward* (en arrière), il s'effectue en parcourant le treillis de la droite vers la gauche.

On définit la variable $\beta(i,t)$ de la façon suivante :

$$\beta(i,t) = P(o_t \ldots o_T | X_t = i)$$

Les variables $\beta(i,t)$ sont calculées en suivant les étapes de l'algorithme décrit ci-dessous.

Algorithme de calcul de $P(o)$ grâce aux probabilités *backward* **:**

1. Initialisation :

$$\beta(i,T) = E(i,o_T),\ 1 \leq i \leq N$$

2. Etape récursive :

$$\beta(i,t) = \sum_{j=1}^{N} \beta(j,t+1)T(i,j)E(i,o_t),\ 1 \leq t \leq T-1,\ 1 \leq i \leq N$$

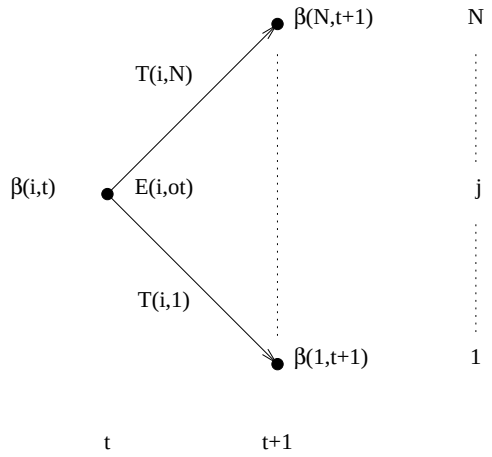

FIG. 1.5 – Calcul de $\beta(i,t)$

3. Calcul de la probabilité totale :

$$P(o) = \sum_{i=1}^{N} \pi(i)\beta(i,1)$$

Combinaison des probabilités *backward* **et** *forward*

Les probabilités forward et backward peuvent être combinées pour calculer $P(o)$ de la façon suivante :

$$P(o) = \sum_{i=1}^{N} \alpha(i,t)\beta(i,t)\ \forall t\ 1 \leq t \leq T$$

Ce résultat est établi en utilisant d'une part la formule des probabilités totales :

$$P(o) = \sum_{i=1}^{N} P(o, X_t = i)$$

puis en remarquant que chacun des termes de la somme peut être exprimée en fonction des probabilités forward et backward de la façon suivante :

$$\begin{aligned} P(o, X_t = i) &= P(o_1, \ldots, o_T, X_t = i) \\ &= P(o_1, \ldots, o_{t-1}, X_t = i, o_t, \ldots, o_T) \\ &= P(o_1, \ldots, o_{t-1}, X_t = i) \times P(o_t, \ldots, o_T | o_1, \ldots, o_{t-1}, X_t = i) \\ &= P(o_1, \ldots, o_{t-1}, X_t = i) \times P(o_t, \ldots, o_T | X_t = i) \\ &= \alpha(i,t)\beta(i,t) \end{aligned}$$

1.2.3 Recherche du chemin le plus probable

Il est souvent intéressant, étant donné un MMC λ et une séquence d'observations $o = o_1 \ldots o_T$ de déterminer la séquence d'états $\hat{x} = \hat{x}_1, \hat{x}_2, \ldots \hat{x}_T$ la plus probable ayant pu générer o.

Une première solution consiste à déterminer toutes les séquences d'états ayant pu générer o, puis de calculer leur probabilités afin de déterminer la plus probable. Il s'agit d'une méthode particulièrement coûteuse car, dans le cas général, il existe N^T chemins possibles. Il existe une méthode beaucoup plus efficace, l'algorithme de Viterbi, qui utilise le treillis vu ci-dessus.

L'idée générale de l'algorithme de Viterbi consiste à déterminer, pour chaque sommet du treillis, le meilleur chemin (le chemin de meilleure probabilité) menant à ce sommet, tout en ayant généré la suite $o_1 \ldots o_t$. On définit pour chaque sommet (j,t) du treillis la variable $\delta(j,t)$ qui est la probabilité du meilleur chemin menant à ce sommet. $\delta(j,t)$ est défini de la façon suivante :

$$\delta(j,t) = \max_{x \in \mathcal{C}_{t-1}} P(x, o_1 \ldots o_t, X_t = j)$$

où $\mathcal{C}_{t-1}$ est l'ensemble des séquences de $t-1$ états de λ et x une de ces séquences.

On définit de plus, pour chaque sommet (j,t) la variable $\psi(j,t)$ dans laquelle est stocké l'état du MMC au temps $t-1$ qui a permis de réaliser le meilleur score, qui n'est donc autre que l'état précédent dans le meilleur chemin menant à (j,t).

L'algorithme lui même ressemble à celui du calcul de la probabilité d'une séquence d'observations vu ci-dessus.

Algorithme de Viterbi

1. Initialisation du treillis :

$$\delta(j,1) = \pi(j)E(j,o_1),\ 1 \leq j \leq N$$

2. Etape récursive :

$$\delta(j,t+1) = \max_{1 \leq i \leq N} \delta(i,t)T(i,j)E(j,o_{t+1}),\ 1 \leq t < T,\ 1 \leq j \leq N$$

stockage du meilleur état précédent :

$$\psi(j,t+1) = \arg\max_{1 \leq i \leq N} \delta(i,t)T(i,j)E(j,o_{t+1}),\ 1 \leq t < T,\ 1 \leq j \leq N$$

3. Détermination du meilleur chemin :

$$\begin{aligned} \hat{x}_T &= \arg\max_{1 \leq i \leq N} \delta(i,T) \\ \hat{x}_t &= \psi(\hat{x}_{t+1}, t+1) \\ P(\hat{x}) &= \max_{1 \leq i \leq M} \delta(i,T) \end{aligned}$$

1.2.4 Estimation des paramètres d'un MMC

Nous avons considéré jusque là que les paramètres d'un MMC (les probabilités initiales, les probabilités de transition et d'émission) étaient connus d'avance. Ce n'est généralement pas le cas et il est nécessaire, avant d'utiliser un MMC, d'estimer la valeur de ses paramètres, tâche que l'on désigne généralement par le terme *apprentissage*.

Nous allons supposer que pour mener à bien cette tâche, nous disposons d'une longue suite d'observations $o = o_1 \ldots o_T$, appelée *données d'apprentissage* qui est sensée être représentative du type de données que le MMC peut produire. Nous allons aussi supposer que la structure du MMC (le nombre d'états et les transitions possibles entre états) est fixée. L'objectif est alors de déterminer les paramètres qui rendent le mieux compte de o, ou, en d'autres termes, de déterminer les paramètres qui, parmi l'ensemble des paramètres possibles, attribuent à o la meilleure probabilité. Si l'on dénote par $P_\lambda(o)$ la probabilité qu'attribue le MMC λ à la suite o, le but de la tâche d'apprentissage est de déterminer le MMC $\hat{\lambda}$ qui maximise $P_\lambda(o)$:

$$\hat{\lambda} = \arg\max_{\lambda} P_\lambda(o)$$

Nous allons supposer dans la suite de cette section que la séquence o a été générée par un MMC. Ceci n'est qu'une vision de l'esprit et l'on ne connaît pas le processus qui est à l'origine de o.

Deux cas peuvent alors se présenter. Dans le premier, appelé apprentissage supervisé, on dispose des données d'apprentissage o et de la séquence d'états $x = x_1 \ldots x_T$ ayant permis la

génération de o. Dans le second cas, appelé apprentissage non supervisé, on ne dispose que de la suite d'observation o.

Apprentissage supervisé

On peut envisager la tâche d'apprentissage supervisé de la façon suivante : Nous avons pu observer un MMC λ générer la séquence d'observables o et, pendant la génération, les états du MMC ne nous étaient pas cachés, de sorte que nous avons pu enregistrer la suite x des états par lesquels λ est passé lors de la génération de o. A l'aide de o et de x, nous voulons déterminer le MMC $\hat{\lambda}$ qui imite le mieux possible λ.
Supposons que λ soit passé par l'état i $\mathcal{C}(i)$ fois lors de la génération de o et que pour un certain nombre de ces passages, noté $\mathcal{C}(a, i)$, ce passage par i a coïncidé avec l'émission du symbole a. Une façon naturelle d'estimer la probabilité d'émission de a lorsque $\hat{\lambda}$ se trouve dans l'état i est de diviser $\mathcal{C}(a, i)$ par $\mathcal{C}(i)$:

$$E_{\hat{\lambda}}(i, a) \approx \frac{\mathcal{C}(a, i)}{\mathcal{C}(i)}$$

De la même façon, si l'on a vu λ passer $\mathcal{C}(i)$ fois par l'état i et que dans un certain nombre de cas, noté $\mathcal{C}(i, j)$, λ a quitté l'état i pour l'état j. Une façon naturelle d'estimer la probabilité de transition de l'état i vers l'état j est de diviser $\mathcal{C}(i, j)$ par $\mathcal{C}(i)$:

$$T_{\hat{\lambda}}(i, j) \approx \frac{\mathcal{C}(i, j)}{\mathcal{C}(i)}$$

Cette méthode d'estimation des probabilités est appelée estimation par maximum de vraisemblance.

Apprentissage non supervisé

Dans le cas d'apprentissage non supervisé, on ne dispose que des données d'apprentissage o et de la structure du MMC $\hat{\lambda}$ dont on veut estimer les probabilités d'émission et de transition. On ne connaît pas de méthode permettant de calculer directement $\hat{\lambda}$, mais il existe une procédure, appelée algorithme de Baum-Welsh ou algorithme forward-backward qui permet de s'en approcher. Il s'agit d'une procédure itérative qui permet de calculer une suite de MMC $\lambda_0, \lambda_1, \ldots, \lambda_n$ où λ_{i+1} est construit à partir de λ_i et tel que :

$$P_{\lambda_{i+1}}(o) \geq P_{\lambda_i}(o)$$

Pour cela, nous allons donner aux paramètres de λ_0 des valeurs arbitraires, qui peuvent être aléatoires, comme elles peuvent être guidées par la connaissance a priori que nous avons du problème.

Etant donné ces paramètres arbitraires, nous allons considérer que o a été généré par λ_0. Cette hypothèse va permettre de calculer la probabilité, notée $\gamma(i, t)$, que λ_0 soit dans l'état i à l'instant t :

$$\begin{aligned}\gamma(i,t) &= P(X_t = i|o) \\ &= \frac{P(X_t = i, o)}{p(o)} \\ &= \frac{\alpha(i,t)\beta(i,t)}{\sum_{j=1}^{N} \alpha(j,t)\beta(j,t)}\end{aligned}$$

A partir de $\gamma(i,t)$, on peut effectuer la somme $\sum_{t=1}^{T} \gamma(i,t)$, qui est la somme des probabilités que λ_0 soit passé par l'état i aux différents instants t de la génération de o. Cette quantité n'est pas facile à interpréter. On pourra remarquer que ce n'est pas une probabilité, pour la simple raison qu'elle peut être supérieure à 1 ; de plus on ne voit à quel évenement une telle probabilité correspond. Nous allons l'interpréter comme une approximation du nombre de fois que λ_0 est passé par l'état i lors de la génération de o.

A ce stade, on se retrouve dans une situation proche de celle de l'apprentissage supervisé car nous disposons d'une estimation du nombre de fois que le MMC est passé par chacun des états, ce qui nous permet de calculer (on dit aussi réestimer) de nouvelles probabilités d'émission, notées E_1, par maximum de vraissemblance :

$$\begin{aligned}E_1(i,a_j) &= \frac{\text{nombre de fois que } \lambda_0 \text{ s'est trouvé dans l'état i et que } a \text{ a été émis}}{\text{nombre de fois que } \lambda_0 \text{ s'est trouvé dans l'état i}} \\ &= \frac{\sum_{t:o_t=a} \gamma(i,t)}{\sum_{t=1}^{T} \gamma(i,t)}\end{aligned}$$

les probabilités intiales peuvent, elles, être réestimées de la façon suivante :

$$\begin{aligned}\pi_1(i) &= \text{probabilité d'être en } i \text{ à l'instant } t = 1 \\ &= \gamma(i,1)\end{aligned}$$

Nous allons suivre une démarche identique pour réestimer les probabilités de transition. Notons $p_t(i,j)$ la probabilité que λ_0 soit passé de l'état i à l'état j entre les instants t et $t+1$ (qu'il soit en i à l'instant t et en j à l'instant $t+1$)[4], voir figure 1.6 :

$$\begin{aligned}p_t(i,j) &= P(X_t = i, X_{t+1} = j|o) \\ &= \frac{P(X_t = i, X_{t+1} = j, o)}{P(o)} \\ &= \frac{\alpha(i,t) \times E(i,o_t) \times T(i,j) \times \beta(j,t+1)}{\sum_{k=1}^{N} \alpha(k,t)\beta(k,t)}\end{aligned}$$

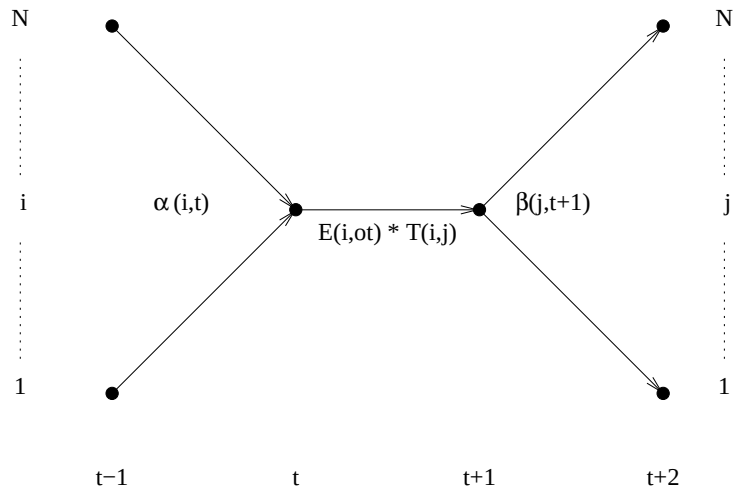

FIG. 1.6 – Calcul de $p_t(i,j)$

On peut maintenant effectuer la somme $\sum_{t=1}^{T} p_t(i,j)$ que nous allons interpréter comme le nombre de fois qu'une transition de i vers j a été empruntée lors de la génération de o et on peut recalculer à partir de cette quantité des nouvelles probabilités de transition T_1 par maximum de vraissemblance :

$$\begin{aligned} T_1(i,j) &= \frac{\text{nombre de fois qu'une transition de i vers j a été empruntée}}{\text{nombre de fois qu'un transition émanant de i a été empruntée}} \\ &= \frac{\sum_{t=1}^{T} p_t(i,j)}{\sum_{t=1}^{T} \gamma(i,t)} \end{aligned}$$

Le nouveau MMC λ_1 dont nous avons calculé les paramètres possède la propriété remarquable d'attribuer à la séquence o une probabilité meilleure ou égale à celle que lui attribuait λ_0 :

$$P_{\lambda_1}(o) \geq P_{\lambda_0}(o)$$

Cette propriété s'explique par le fait que lors du calcul des paramètres de λ_1, par maximum de vraissemblance, nous avons augmenté la probabilité des transitions et des émissions qui étaient à l'origine de la génération de o, et ce faisant, diminué les autres probabilités.

Ainsi, en réitérant le pocessus de réestimation des probabilités, nous obtiendrons des paramètres attribuant une probabilité de plus en plus élevée à la séquence o, jusqu'à ce qu'une valeur limite soit atteinte, pour un MMC λ_n. Ce MMC n'est cependant pas le meilleur possible (celui qui attribue à o la meilleure probabilité possible), il peut s'agir d'un maximum local, qui dépend des paramètres initiaux desquels nous sommes partis, comme l'illustre schématiquement la figure 1.7.

[4]On pourra noter que $\gamma(i,t) = \sum_{j=1}^{N} p_t(i,j)$

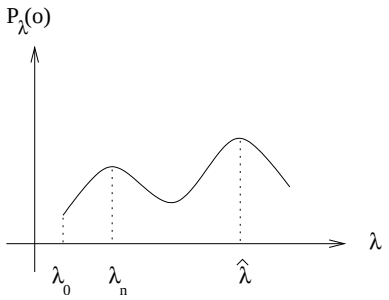

FIG. 1.7 – Maximum local : en partant de λ_0, on n'atteindra pas $\hat{\lambda}$ par l'algorithme de Baum-Welsh.