

DIIC - INC, 3^e année

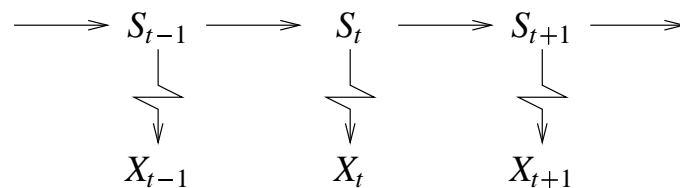
Module RAPP

TP 3 - Estimation bayésienne

Hervé JÉGOU, Guy LORETTE

8 février 2006

Lors de ce TP, vous allez estimer les états d'une chaîne de Markov cachée, ce qui vous permettra normalement de déduire la provenance des séquences que vous avez manipulées depuis le premier TP. La chaîne de Markov considérée est illustrée ci-dessous.



Il s'agit de retrouver une séquence $\mathbf{S} = S_1, \dots, S_n$ à partir de son observation bruitée $\mathbf{X} = X_1, \dots, X_n$. Les séquences que vous avez manipulées peuvent être considérées comme des versions bruitées de la séquence originales. La matrice des probabilités de transition que vous avez calculée [*sic*] lors du TP 2 correspond aux probabilités $\mathbb{P}(X_t|X_{t-1})$, qui pourraient être déduites à partir des probabilités de la source $\mathbb{P}(S_t|S_{t-1})$.

Les séquences bruitées ont été obtenues à partir de textes en appliquant la fonction $f : \mathcal{C} \rightarrow \mathcal{X}$ définie ci-dessous, où \mathcal{C} désigne l'espace des caractères et $\mathcal{X} = \{0, \dots, 9\}$.

$$f : s \mapsto \begin{cases} 0 & \text{si } s \text{ est un caractère de ponctuation ou un espace} \\ 1 & \text{si } s \in \{a,j,s\} \\ 2 & \text{si } s \in \{b,k,t\} \\ 3 & \text{si } s \in \{c,l,u\} \\ 4 & \text{si } s \in \{d,m,v\} \\ 5 & \text{si } s \in \{e,n,w\} \\ 6 & \text{si } s \in \{f,o,x\} \\ 7 & \text{si } s \in \{g,p,y\} \\ 8 & \text{si } s \in \{h,q,z\} \\ 9 & \text{si } s \in \{i,r\} \end{cases}$$

Les séquences que vous avez manipulées correspondent à deux types de textes, respectivement de langue française et anglaise. Dans ce TP, elles seront modélisées par des modèles

de source markovienne, qui vous sont fournis dans le fichier de paramètres `tp3.param` sous la forme de matrices de probabilités de transition $\mathbb{P}(S_t|S_{t-1})$. Ces modèles ont été générés à partir des données originales. Remarquez que, pour être parfaitement juste, il faudrait utiliser des probabilités de transition de la langue anglaise et française apprises sur un corpus de taille importante. De telles probabilités vous sont fournies sous forme logarithmique pour l'anglais dans le fichier `proba_anglais`.

Brainstorming

L'estimation de la séquence \mathbf{S} peut se faire avec l'un des deux estimateurs suivants.

- *Le maximum a posteriori* : cet estimateur renvoie la séquence qui a la plus grande probabilité d'avoir été émise sachant les observations $\hat{S}_1, \dots, \hat{S}_k$:

$$\tilde{S} = \arg \max_{(S_1, \dots, S_n) \in \mathcal{C}^*} \mathbb{P}(S_1, \dots, S_n | X_1, \dots, X_n). \quad (1)$$

Cet estimateur maximise la probabilité que la séquence choisie soit correcte. Il est usuellement implémenté en utilisant l'algorithme de Viterbi. Cet algorithme peut être vu comme un cas particulier de l'algorithme de Dijkstra. C'est un algorithme de programmation dynamique. En cela, il s'apparente au calcul de la distance de Levenshtein (les probabilités en plus).

- *Le maximum des marginales postérieures* : cet estimateur renvoie la séquence composée, pour chaque instant t , du symbole X_t qui a plus grande probabilité d'avoir été émis, i.e.

$$\forall t, \tilde{S}_t = \arg \max_{\mathcal{C}} \mathbb{P}(S_t | X_1, \dots, X_n). \quad (2)$$

Cet estimateur maximise le nombre de symboles correctement décodés. *** Il minimise donc le taux d'erreur symbole de la séquence. *** L'algorithme qui permet de le calculer de manière optimale est l'algorithme BCJR. Précisons que cet algorithme se rencontre sous d'autres noms, selon la communauté scientifique considérée : algorithme forward/backward, algorithme somme/produit.

Dans un premier temps, vous allez participer à une séance de *brainstorming* afin de vous remémorer le principe de l'algorithme de Viterbi. Je vous invite à une collaboration active.

Estimation d'une chaîne de Markov cachée : algorithme de Viterbi

Question 1. Comment définissez-vous la probabilité $\mathbb{P}(X_t|S_t)$?

Question 2. Pour chaque instant t , combien de probabilités (vraisemblances) de chemins devez-vous calculer ?

Question 3. Implantez l'algorithme de Viterbi et appliquez-le aux séquences $(D_i)_i$ pour les deux modèles (français et anglais).

Question 4. Les probabilités a posteriori des chemins renvoyées par l'algorithme permettent-elles de classifier correctement les séquences $(D_i)_i$?

Question 5. *L'algorithme de Viterbi permet-il de retrouver les séquences originales si les versions bruitées qui lui sont fournies sont les séquences $(N_j)_j$? Commentez.*

L'algorithme de Viterbi permet de calculer la séquence la plus probable, mais celle-ci n'est pas toujours la plus pertinente. Il est ainsi souvent préférable de trouver la séquence dont les éléments sont marginalement les plus probables.

Question 6. *Implantez l'algorithme backward/forward pour trouver la séquence qui maximise les marginales postérieures. Que pensez-vous du résultat obtenu ?*

Question 7. *Utiliser les probabilités de transition de la langue anglaise en remplacement des probabilités fournies dans le fichier tp3.param. Commentez.*