

DIIC - INC, 3^e année

Module RAPP

TP 2 - Distances et introduction aux processus markoviens

Hervé JÉGOU, Guy LORETTE

31 janvier 2006

LORS de ce TP, vous allez vous familiariser avec les fonctions de distance entre séquences, puis manipuler de manière très élémentaire des processus bayésiens. Les techniques d'estimation bayésienne proprement dites seront considérées dans le dernier TP.

Dans un premier temps, vous allez considérer des distances qui permettent de comparer des séquences. Vous effectuerez ensuite un apprentissage simple d'un modèle markovien et calculerez les probabilités *a posteriori* des séquences considérées. Enfin, vous implanterez un petit jeu qui devrait vous permettre d'entrevoir la difficulté qu'il peut y avoir pour un humain à produire des séquences aléatoires.

Comme pour le TP1, vous rédigerez votre compte-rendu au fur et à mesure du TP et vous le rendrez à la fin des quatre heures.

Questions de distance

Dans cette partie, vous allez considérer les nouvelles séquences N_1 à N_6 . Le but est de retrouver quelles sont les séquences qui sont les plus proches, car l'on sait que ces séquences ont été obtenues par modification des séquences D_1 à D_{29} . Vous allez donc être amenés à implanter des fonctions de distance.

Question 1. *Implantez une distance de Hamming entre deux séquences. À quelle(s) limitation(s) êtes-vous confrontés pour effectuer la comparaison de deux séquences ? Avez-vous des idées pour y remédier ?*

Question 2. *Écrivez une fonction qui implante la distance de Levenshtein entre deux séquences. Comparez les séquences $(N_j)_j$ aux séquences $(D_i)_i$. Comparez également les séquences $(N_j)_j$ entre elles. Commentez.*

Dans la suite, vous considérerez également version normalisée de la distance de Levenshtein. Pour cela, vous diviserez la distance précédente par la longueur de la première des séquences considérées. Remarquez que cette quantité n'est alors plus symétrique.

Question 3. Modifiez votre fonction de la distance de Levenshtein pour qu'elle renvoie, en plus de la distance proprement dite, les modifications effectuées pour transformer une séquence en une autre. Précisez éventuellement vos choix.

Question 4. Quelles limites entrevoyez-vous pour la distance de Levenshtein ?

Question 5. Résumez dans un tableau la correspondance entre les séquences $(D_i)_i$ et $(N_j)_j$ en précisant si possible quelles sont les altérations que les premières ont subi pour obtenir les secondes.

Modélisation markovienne

On rappelle que les séquences proposées (séquences D_1 à D_{29}) ont été classifiées lors du premier TP en deux classes distinctes, respectivement indicées par les entiers pairs et impairs.

Question 6. Écrivez une fonction qui calcule la matrice des probabilités de transition à partir d'un sous-ensemble de séquences repérées par leur indices. Commentez.

Question 7. Déduisez-en la matrice des probabilités de transition de chacune des classes.

Question 8. Écrivez une fonction qui prend en entrée

- une matrice de probabilités de transition définissant une chaîne de Markov,
- une séquence,

et qui renvoie la probabilité a posteriori que la chaîne de Markov passée en paramètre ait généré ladite séquence.

Question 9. Calculez alors la probabilité qu'une séquence ait été générée par les processus de Markov associés à chacune des classes. Les résultats corroborent-ils ceux obtenus lors du premier TP ? Que faudrait-il faire pour être plus juste ?

Jeu

Le but de cette partie est d'écrire un petit jeu. Le principe est le suivant :

- il y a deux "participants" : le programme et un joueur humain ;
- le joueur choisit 0 ou 1, en essayant de "tromper" le programme ;
- le programme prédit le choix de l'humain sans utiliser l'information entrée par l'humain ;
- le programme vérifie alors si sa prédiction est correcte. Si oui, il gagne un point, sinon c'est l'humain qui marque un point.

Le programme sera implanté en modélisant le comportement de l'humain comme un processus de Markov d'ordre m . Il sera donc modélisé par une chaîne de Markov à 2^m états. Initialement, les probabilités de transition de cette chaîne sont supposées uniformes. Elles sont ensuite modifiées en fonction des choix du joueur humain. À chaque tour de jeu, le programme calcule la probabilité du 0 et du 1. Il y a deux manières de procéder au choix effectué par le programme :

- le programme choisit la réalisation (0 ou 1) la plus probable ;
- le programme tire au hasard selon la probabilité calculée.

Question 10. *Implantez le jeu.*

Question 11. *Jouez !*