Analyse de données et modèles stochastiques

10 novembre 2023

Durée : 1h30

> Supports et notes de cours autorisées. Calculatrices autorisées.
>
> Les exercices sont indépendants. Les questions au sein des exercices peuvent pour la plupart être considérées comme indépendantes. Le nombre de point par exercice est donné à titre indicatif et pourra être modifié a posteriori, dans les limites du raisonnable.
>
> Quand vous répondez aux questions, ne vous limitez pas à un simple résultat. Donnez le détail de vos calculs et précisez le raisonnement suivi.

Exercice 1 – Chaînes de Markov cachées (10 points)

> On utilisera pour cet exercice la table des logarithmes simplifiés donnée en annexe.

On s'intéresse à l'état de forme d'un·e athlete au cours du temps. Iel peut être en repos (R), en reprise d'activité (A), en forme (F), ou en super forme (S). Les mesures commencent toujours alors que l'athlète est en reprise d'activite (A) et les probabilitées de passer d'un état à un autre sont les suivantes :

	vers R	*vers* A	*vers* F	*vers* S
R	0.3	0.7	0.0	0.0
A	0.0	0.3	0.7	0.0
F	0.0	0.0	0.8	0.2
S	0.9	0.0	0.0	0.1

.

Questions

1. Dessiner l'automate associé.
2. Quelle est la probabilité de la séquence d'états $A\ F\ S\ S\ S$?

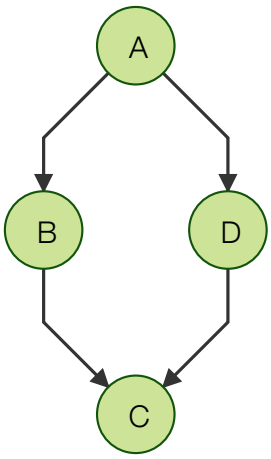
Lors des entrainement, les temps de l'athlète peuvent être de niveau a, b, ou c, et dépendent de l'état de forme come suit :

$$P(temps|forme) = \begin{array}{c|ccc} & vers\ a & vers\ b & vers\ c \\ R & 0.9 & 0.1 & 0.0 \\ A & 0.2 & 0.8 & 0.0 \\ F & 0.1 & 0.5 & 0.4 \\ S & 0.1 & 0.3 & 0.6 \end{array}.$$

Questions

- 3. À l'aide de l'algorithme de Viterbi, donnez la séquence d'état la plus vraisemblable pour la séquence d'observations $a\ b\ b\ c\ a$?
- 4. Certains états de forme ont des durées typiques (par exemple pas plus de quelques semaines pour S) ce qui n'est pas pris en compate par la modélisation actuelle. Expliquez en quoi une modélisation par champs aléatoires conditionnels (CRF) permettrait d'améliorer ce fait.

Exercice 2 – Réseaux bayésiens (6 points)



On considère le réseau bayésien ci-dessus. Les variables sont binaires et le modèle est défini par les paramètres suivants :

$$P[A=0]=0.3 \quad P[B=0\mid A]=\begin{array}{|c|c|}\hline A & \\ \hline 0 & 0.5 \\ 1 & 0.7 \\ \hline \end{array} \quad P[D=0\mid A]=\begin{array}{|c|c|}\hline A & \\ \hline 0 & 0.2 \\ 1 & 0.6 \\ \hline \end{array}$$

$$P[C=0\mid B,D]=\begin{array}{|cc|c|}\hline B & D & \\ \hline 0 & 0 & 0.3 \\ 0 & 1 & 0.1 \\ 1 & 0 & 0.8 \\ 1 & 1 & 0.5 \\ \hline \end{array}.$$

Questions

1. Donnez la forme factorisée de la probabilité conjointe $P[A, B, C, D]$.
2. Dessinez le modèle à base de facteurs correspondant.
3. Calculez $P[C = 1|A = 0, D = 0]$.
4. Indiquez deux variables qui sont indépendantes conditionnellement à une autre variable et prouvez cette indépendance en vous appuyant sur la factorisation de la probabilité conjointe $P[A, B, C, D]$.

Exercice 3 – Test d'hypothèse (4 points)

Vous souhaitez comparer votre algorithme (Our) à celui de l'état de l'art (SOTA) en ce qui concerne le temps de calcul. Vous les exécutez dans 5 contextes différents et obtenez les temps de calcul (en secondes) :

Setting	$S1$	$S2$	$S3$	$S4$	$S5$
`Our`	1	1	3	2	3
`SOTA`	3	1	2	3	6

Questions

1. Le temps de calcul de Our est-il différent du temps de calcul de SOTA avec un niveau de confiance de 5% ?

On rappelle que pour une variable aléatoire X suivant une loi de Student à k degrés de liberté, le quantile d'ordre 0.95 est :

- 2.353 pour $k = 3$,
- 2.132 pour $k = 4$,
- 2.015 pour $k = 5$,
- 1.943 pour $k = 6$,
- 1.895 pour $k = 7$.

Annexe

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$ln(x)$	-2.3	-1.6	-1.2	-0.9	-0.7	-0.5	-0.4	-0.2	-0.1