Aix*Marseille université Initiative d'excellence **Master SID**

Raisonnement dans l'incertain

TD n°4 : Apprentissage de structure de réseau bayésien

Exercice 1 – Apprentissage par PC

Soit trois variables aléatoires booléennes A, B, C. On a constitué la base de données suivante :

A	B	C
a_1	b_1	c_1
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_1	c_2
a_2	b_1	c_1
a_2	b_1	c_2
a_1	b_2	c_1
a_1	b_2	c_2
a_2	b_2	c_1
a_2	b_2	c_1
a_2	b_2	c_2
a_2	b_2	c_2

Dans cet exercice, on suppose que toute probabilité conditionnelle portant sur les variables A, B, C se déduit par normalisation des fréquences observées dans la base, c'est-à-dire qu'en multipliant ces fréquences par une constante de telle sorte qu'elles somment à 1, on obtient des probabilités. Par exemple, on observe 6 instances de a_1 et 6 instances de a_2. Par conséquent, $P(A = a_1) = P(A = a_2) = 6 \times k$, avec la constante $k = 1/12$ afin d'obtenir $P(A = a_1) + P(A = a_2) = 1$.

Q 1.1 En utilisant des tests d'indépendance conditionnels fondés sur les probabilités, appliquez l'algorithme PC pour apprendre le squelette du réseau bayésien ayant généré cette base.

Q 1.2 Appliquez les règles R1, R2, R3 de PC. Quel CPDAG obtient-on ?

Q 1.3 Finalisez les orientations en respectant l'ordre topologique $A \precsim B \precsim C$.

Q 1.4 Calculez par maximum de vraisemblance les paramètres du réseau bayésien obtenu à la question précédente.

Exercice 2 – Apprentissage par PC – bis

Soit trois variables aléatoires booléennes A, B, C dont on a observé les occurrences ci-dessous :

A	B	C
a_1	b_1	c_1
a_2	b_2	c_1
a_2	b_1	c_2
a_2	b_2	c_2
a_2	b_2	c_1
a_1	b_1	c_1
a_2	b_2	c_1
a_2	b_1	c_1
a_1	b_2	c_2

On suppose ici que toute probabilité conditionnelle portant sur les variables A, B, C se déduit par normalisation des fréquences observées dans la base, c'est-à-dire qu'en multipliant ces fréquences par une constante de telle sorte qu'elles somment à 1, on obtient des probabilités. Par exemple, on observe 3 instances de a_1 et 6 instances de a_2. Par conséquent, $P(A = a_1) = 3 \times k$ et $P(A = a_2) = 6 \times k$, avec la constante $k = 1/9$ afin d'obtenir $P(A = a_1) + P(A = a_2) = 1$.

Q 2.1 En utilisant des tests d'indépenance conditionnels fondés sur les probabilités, appliquez l'algorithme PC pour apprendre le squelette du réseau bayésien ayant généré cette base.

Q 2.2 Appliquez les règles R1, R2, R3 de PC. Quel CPDAG obtient-on ?

Q 2.3 Déterminez un réseau bayésien compatible avec le CPDAG trouvé dans la question précédente. Estimez par maximum de vraisemblance les tables de probabilité conditionnelles des nœuds du réseau.

Exercice 3 – Apprentissage par PC – ter

Soit trois variables aléatoires booléennes A, B, C dont on a observé les occurrences suivantes :

A	B	C
a_1	b_1	c_1
a_1	b_1	c_1
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_2	c_1
a_1	b_2	c_2
a_2	b_1	c_2
a_2	b_1	c_2
a_2	b_1	c_2
a_2	b_1	c_1
a_2	b_2	c_2
a_2	b_2	c_2

On suppose ici que toute probabilité conditionnelle portant sur les variables A, B, C se déduit par normalisation des fréquences observées dans la base, c'est-à-dire qu'en multipliant ces fréquences par une constante de telle sorte qu'elles somment à 1, on obtient des probabilités. Par exemple, on observe 6 instances de a_1 et 6 instances de a_2. Par conséquent, $P(A = a_1) = P(A = a_2) = 6 \times k$, avec la constante $k = 1/12$ afin d'obtenir $P(A = a_1) + P(A = a_2) = 1$.

Q 3.1 En utilisant des tests d'indépenance conditionnels fondés sur les probabilités, appliquez l'algorithme PC pour apprendre le squelette du réseau bayésien ayant généré cette base.

Q 3.2 Appliquez les règles R1, R2, R3 de PC. Quel CPDAG obtient-on ?

Exercice 4 – Test d'indépendance du χ^2

Un échantillon de 200 contribuables est prélevé afin de vérifier si le revenu brut annuel d'un individu est un caractère dépendant du niveau de scolarité de l'individu. Les observations recueuillies sont données dans le tableau suivant :

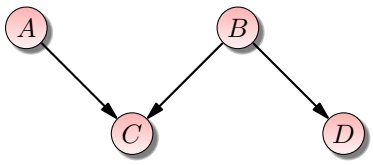
scolarité (années) → revenu (kF) ↓	$[0;7[$	$[7;12[$	$[12;14[$	$[14;\rightarrow[$	total
$[0;75[$	17	14	9	5	45
$[75;120[$	12	37	11	5	65
$[120;200[$	7	20	20	8	55
$[200;\rightarrow[$	4	9	10	12	35
total	40	80	50	30	200

Q 4.1 On admet que les fréquences relatives déduites des marges du tableau donnent les vraies lois de probabilité, $P(R)$ et $P(S)$ des variables $R(evenu)$ et $S(colarité)$. Donner le tableau des fréquences théoriques, $200 \times P(R, S)$, correspondantes en cas d'indépendance des deux variables.

Q 4.2 Calculer le χ^2. Expliquez pourquoi il y a 9 degrés de liberté. Doit-on rejeter l'hypothèse d'indépendance au risque $\alpha = 5\%$?

Exercice 5 – Score AIC

On considère le réseau bayésien ci-dessous, dans lequel chacune des variables aléatoires a 2 modalités (variables booléennes).



Calculez le score AIC de ce réseau bayésien avec la base de données suivante :

A	B	C	D
0	1	1	0
1	0	0	1
0	1	1	1
1	0	1	0

Exercice 6 – Algorithme K2

Une entreprise est répartie sur 4 sites $S_1, \ldots, S_4$. Chaque site possède des panneaux solaires produisant une partie de sa consommation électrique. Des capteurs $C_1, \ldots, C_4$ sont installés dans ces sites, qui indiquent si, à un moment donné, les panneaux solaires permettent au site d'être autonôme en énergie ($C_i = 1$) ou bien si un complément provenant d'un fournisseur d'énergie est nécessaire ($C_i = 0$). Le tableau ci-dessous montre les données recueillies par les capteurs :

C_1	C_2	C_3	C_4
0	1	1	0
1	1	0	1
0	1	0	1
1	0	1	0

Q 6.1 Appliquez l'algorithme K2 afin de déterminer la structure du réseau bayésien le plus vraisemblable. Pour cela, vous utiliserez le score K2 ainsi que l'ordre topologique $C_1 \precsim C_2 \precsim C_3 \precsim C_4$.