Université Mohamed Khider-Biskra
Faculté des sciences exactes et des sciences de la nature et de la vie
Département d'informatique

M2 SIOD — Analyse des données séquentielles
14 Jan 2019 — 10:00-11:30, S5,S6

Examen

Barème: 1: 3 pts / **2:** 3 pts / **3:** 3 pts / **4:** 1.5 pt / **5:** 4 pts / **6:** 5.5 pts

Nous souhaitons analyser la progression des étudiants du département d'informatique pour une période de dix années de 2009/2010 jusqu'à 2018/2019. Nous avons rassemblé pour cet objectif des informations représentent les états de 07 étudiants de master inscrits en première année pour l'année 2009/2010. Les états considérés sont les suivants :

- N : Inscrit normalement
- R : Redoublant
- C : Inscrit avec crédit
- D : Diplômé
- E : Exclu
- A : Abandon
- B : Bloqué

La base utilisée est représentée dans le tableau suivant :

Année \ Etudiant	E_1	E_2	E_3	E_4	E_5	E_6	E_7
2009/2010	N	N	N	N	N	N	N
2010/2011	N	N	R	N	C	N	N
2011/2012	R	N	R	B	R	N	A
2012/2013	C	N	C	N	R	N	A
2013/2014	R	N	R	C	A	B	A
2014/2015	N	D	E	N	A	N	A
2015/2016	R	D	E	N	A	C	A
2016/2017	C	D	E	D	A	R	A
2017/2018	D	D	E	D	A	D	A
2018/2019	D	D	E	D	A	D	A

Questions :

1. Nous voulons comparer les parcours de deux étudiants :
   (a) Parmi les mesures de similarité des séquences de données : sac de caractère, p-spectrum, LCP et LCS, laquelle qui convient ?
   (b) Justifier votre réponse.
   (c) Démontrer sur un exemple de la base.

2. Dessiner pour cette base :
   - Le D-plot
   - La séquence des états modaux
3. Comparer la turbulence des deux étudiants E_4 et E_6.
4. Calculer la matrice des taux de transition de l'étudiant E_1.
5. Sachant que les étudiants E_2 et E_4 ont poursuivi leur étude en doctorat et que les étudiants E_3, E_5, E_6 et E_7 ne l'ont pas fait, et en utilisant la méthode 3PPV avec la distance bag of characters, dites si l'étudiant E_1 poursuivra ses études en doctorat ou non.
6. En utilisant l'algorithme AprioriAll avec un support minimum de 50 %, trouver les motifs fréquents séquentiels.

Bonne Chance

Dr A.Djeffal

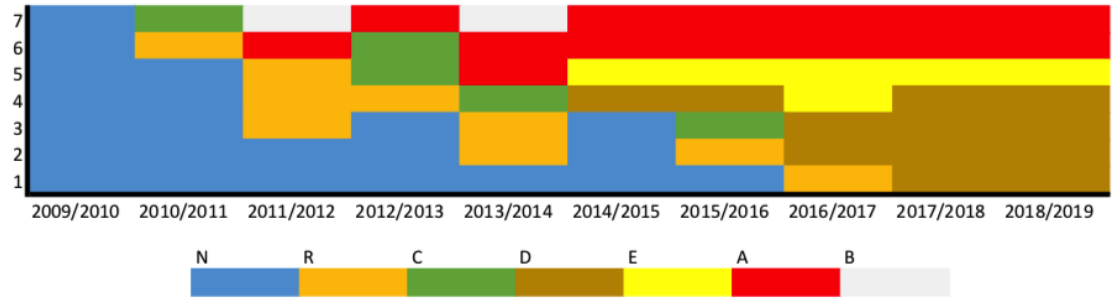
Corrigé type

1. Nous voulons comparer les parcours de deux étudiants :
   (a) La mesure qui convient est le "sac de caractères" **(1 pt)**
   (b) Puisque il s'agit des parcours des étudiants, ce qui compte c'est le nombre d'états et non pas leur ordre : plus il est similaire entre les parcours de deux étudiants, plus les parcours sont proches. Donc la mesure la plus adéquate est celle de sac de caractère. **(1 pt)**
   (c) Prenons l'exemple E_1 avec les deux exemples E_4 et E_7, on voit que le parcours de E_1 ressemble beaucoup à celui de E_4 et très différent de celui de E_7

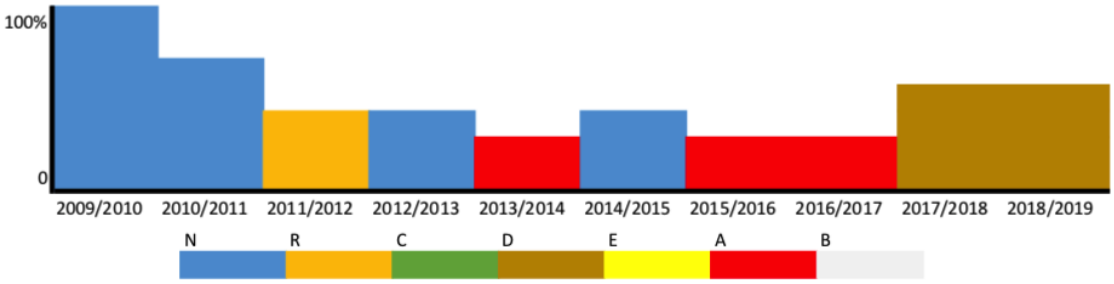
	Sac de caractères	LCS	LCP	2-spectrum
Similarité(E1,E4)	23	6 (NNCNDD)	2 (NN)	4
Similarité(E1,E7)	6	2(NN)	2 (NN)	1

   La similarité la plus importante est celle du sac de caractères. **(1 pt)**

2. Dessins :
   – Le D-plot **(1.5 pt)**



   – La séquence des états modaux **(1.5 pt)**



3. Comparaison des turbulences des deux étudiants E_4 et E_6.
   (a) Turbulence de E_4=NNBNCNNDDD
   – Nombre de Sous séquences $\Phi(x) = 198$
   – Durées passées dans les états : 2,1,1,1,2,3
   S_t^2 = Variance = 0.55

- $\overline{(t)=10/4=2.5}$
  $S^2_{t,max} = (10-1)(1-t)^2 = 20.25$
- $T(x) = log_2(\Phi(x)\frac{S^2_{t,max}(x)+1}{S^2_t+1}) = log_2(198 \times \frac{20.25+1}{0.55+1}) = 12.01$ **(1.25 pt)**

(b) Turbulence de E_6=NNNNBNCRDD

- Nombre de Sous séquences $\Phi(x) = 197$
- Durées passées dans les états : 4,1,1,1,1,2
  S^2_t = Variance = 1.22
- $\overline{(t)=10/5=2}$
  $S^2_{t,max} = (10-1)(1-t)^2 = 9$
- $T(x) = log_2(\Phi(x)\frac{S^2_{t,max}(x)+1}{S^2_t+1}) = log_2(197 \times \frac{9+1}{1.22+1}) = 9.79$ **(1.25 pt)**

(c) La séquence de E_6 est moins turbulente que celle de E_4 **(0.25 pt)**

4. Matrice des taux de transition de l'étudiant E_1. **(1.5 pt)**

	N	R	C	D	E	A	B
N	1/3	2/3	0	0	0	0	0
R	1/3	0	2/3	0	0	0	0
C	0	1/2	0	1/2	0	0	0
D	0	0	0	1	0	0	0
E	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0

5. Les étudiants E_2 et E_4 ont poursuivi leur étude en doctorat et que les étudiants E_3, E_5, E_6 et E_7 ne l'ont pas fait, on utilisant la méthode 3PPV avec la distance bag of charactères :
   - Réprésentation en sac de caractères **(1.5 pt)**

	N	R	C	D	E	A	B	Classe
E_1	3	3	2	2	0	0	0	?
E_2	5	0	0	5	0	0	0	Oui
E_3	1	3	1	0	5	0	0	Non
E_4	5	0	1	3	0	0	1	Oui
E_5	1	2	1	0	0	6	0	Non
E_6	5	1	1	2	0	0	1	Non
E_7	2	0	0	0	0	8	0	Non

   - Distances **(1.5 pt)**

$D(E_1, E_2)$	$\frac{25}{\sqrt{26\times50}} = 0.69$
$D(E_1, E_3)$	$\frac{14}{\sqrt{26\times36}} = 0.45$
$D(E_1, E_4)$	$\frac{23}{\sqrt{26\times36}} = 0.75$
$D(E_1, E_5)$	$\frac{11}{\sqrt{26\times42}} = 0.33$
$D(E_1, E_6)$	$\frac{24}{\sqrt{26\times32}} = 0.83$
$D(E_1, E_7)$	$\frac{6}{\sqrt{26\times68}} = 0.14$

   - Tri des étudiants selon la distance de E_1 **(0.5 pt)**

	Etudiant	Distance	Classe
1	$D(E_1, E_6)$	0.83	Non
2	$D(E_1, E_4)$	0.75	Oui
3	$D(E_1, E_2)$	0.69	Oui
4	$D(E_1, E_3)$	0.45	Non
5	$D(E_1, E_5)$	0.33	Non
6	$D(E_1, E_7)$	0.14	Non

- Les trois plus proches voisins : $E_6, E_4, E_2 \Rightarrow$ La classe de E_1 est **"Oui"** **(0.5 pt)**

6. En utilisant l'algorithme AprioriAll avec un support minimum de 50 %, trouver les motifs fréquents séquentiels.

   - Support min = 50% $\Rightarrow$ Frésuance min = $\frac{7}{2} = 3.5 \approx 4$
   - Motifs de longueur 1 = $\{N(7), R(4), C(5), D(4), E(1), A(2), B(2)\}$
   - $F_1 = \{N, R, C, D\}$ **(1 pt)**
   - Pas besoin de faire le mapage puisque les motifs de longueur 1 sont simples
   - Candidats de longueur 2

Motifs de longueur 2	Fréquence	Motifs de longueur 2	Fréquence
NN	5	CN	2
NR	4	CR	4
NC	5	CC	1
ND	4	CD	3
RN	1	DN	0
RR	3	DR	0
RC	2	DC	0
RD	2	DD	4

   - $F_2 = \{NN, NR, NC, ND, CR, DD\}$ **(1 pt)**
   - Candidats de longueur 3

Motif L2	Jointure	Elagage	Fréquence	Motif L2	Jointure	Elagage	Fréquence
NN	NNN	NNN	4	ND	NDN	X	
	NNR	NNR	2		NDR	X	
	NNC	NNC	3		NDC	X	
	NND	NND	4		NDD	4	
NR	NRN	X		CR	CRR	X	
	NRR	X		DD	DDD	2	
	NRC	X					
	NRD	X					
NC	NCN	X					
	NCR	X					
	NCC	X					
	NCD	X					

   - $F_3 = \{NNN, NND, NDD\}$ **(1.5 pt)**
   - Candidats de longueur 4

Motif L3	Jointure	Elagage	Fréquence
NNN	NNNN	NNNN	4
	NNND	NNND	4
NND	NNDN	X	
	NNDD	4	
NDD	NDDD	4	

- $F_4 = \{NNNN, NNND, NNDD\}$ **(1 pt)**
- Candidats de longueur 5

Motifs L4	Jointure	Elagage	Fréquence
NNNN	NNNNN	NNNNN	3
	NNNND	NNNND	3
NNND	NNNDN	X	
	NNNDD	X	
NNDD	NNDDD	X	

- $F_5 = \{\phi\}$ **(1 pt)**