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
13 Janvier 2024 — 12:30-14:00

Examen

Question de cours (4 pts: 1,5 + 1,5 + 1)

1. Quelles sont les méthodes utilisées pour détecter les anomalies dans une série temporelle ?
2. Expliquer pourquoi la tâche de détection de motifs séquentiels dans la fouille de données séquentielle est un problème difficile.
3. Quelle est la solution pour faire face à cette difficulté ?

Exercice 1 (6 pts: 1 + 2 + 2 + 1)

On considère une série de ventes d'une entreprise en milliers d'euros entre 1970 et 1978.

t_i	1970	1971	1972	1973	1974	1975	1976	1977	1978
y_i	32	38	48	52	61	73	80	84	95

1. Représenter graphiquement cette série. Commenter.
2. Calculer la distance entre les deux sous-séquences représentant respectivement les quatres premières années et les quatres dernières années de la série en utilisant la distance DTW. Que déduisez vous ?
3. On se propose d'ajuster à cette série une tendance linéaire de la forme $f(t) = a\,t + b$. Déterminer a et b par la méthode des moindres carrés.
4. Proposer une prévision des ventes pour les deux années qui suivent.

Exercice 2 (10 pts: 2 + 2 + 4 +2)

Soit la base de données séquentielle suivante représentant l'historique des achats des clients.

ID séquence	Séquence
1	< (1,5) (2) (3) (4) >
2	< (1) (3) (4) (3,5) >
3	< (1) (2) (3) (4) >
4	< (1) (3) (5) >
5	< (4) (5) >

1. Calculer et comparer les turbulences des séquences d'achats des clients trois et quatre.
2. Générer une seule matrice des taux de transitions comportant les probabilités de transition à une position donnée d'un état à l'autre pour toutes les séquences du tableau.
3. Appliquer l'algorithme GSP à l'ensemble des données du tableau en utilisant un support minimum s = 33 % pour déterminer toutes les séquences fréquentes.
4. En déduire les motifs fréquents maximums.

Bonne chance!

Dr D.AKROUR

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
13 Janvier 2024 — 12:30-14:00

Examen (corrigé type)

Question de cours (4 pts: 1,5 + 1,5 + 1)

1. Quelles sont les méthodes utilisées pour détecter les anomalies dans une série temporelle ?
2. Expliquer pourquoi la tâche de détection de motifs séquentiels dans la fouille de données séquentielle est un problème difficile.
3. Quelle est la solution pour faire face à cette difficulté ?

(voir le cours)

Exercice 1 (6 pts: 1 + 2 + 2 + 1)

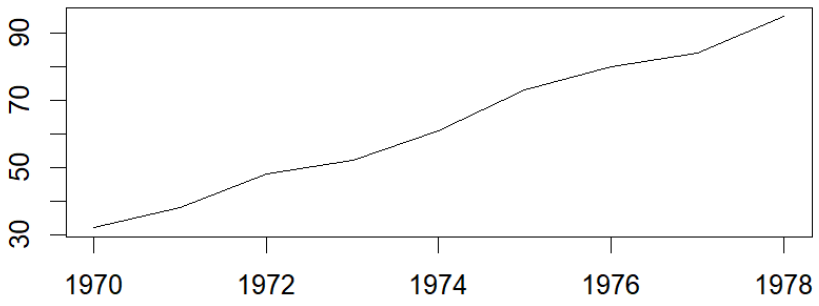
On considère une série de ventes d'une entreprise en milliers d'euros entre 1970 et 1978.

t_i	1970	1971	1972	1973	1974	1975	1976	1977	1978
y_i	32	38	48	52	61	73	80	84	95

1. Représenter graphiquement cette série. Commenter.

réponse:

(0.5pt)



courbe qui a une tendance croissante (0.5pt)

2. Calculer la distance entre les deux sous-séquences représentant respectivement les quatres premières années et les quatres dernières années de la série en utilisant la distance DTW. Que déduisez vous ?

   réponse:

   DTW

95	204	186	166	162
84	141	129	119	140
80	89	83	108	129
73	41	76	101	122
	32	**38**	**48**	**52**

   distance= 1/162 = 0,006172 (2pt)

3. On se propose d'ajuster à cette série une tendance linéaire de la forme $f(t) = a\,t + b$. Déterminer a et b par la méthode des moindres carrés.

   réponse:

   moy(t)=1974 **moy(y)**=62.5556 **cov(t, x)**=52.7778 **var(t)**=6.6666667 (1pt)

   a = 7,9166696 **b** = -15564,95 (1pt)

4. Proposer une prévision des ventes pour les deux années qui suivent.

   réponse:

   pour l'année 1979 y = 102,13 (0,5pt)

   pour l'année 1980 y = 110,05 (0,5pt)

Exercice 2 (10 pts: 2 + 2 + 4 +2)

Soit la base de données séquentielle suivante représentant l'historique des achats des clients.

ID séquence	Séquence
1	< (1,5) (2) (3) (4) >
2	< (1) (3) (4) (3,5) >
3	< (1) (2) (3) (4) >
4	< (1) (3) (5) >
5	< (4) (5) >

1. Calculer et comparer les turbulences des séquences d'achats des clients trois et quatre.

   réponse: (2pt)

   séquence 3: 4
   séquence 4: 3
   la première séquence est plus turbulente

2. Générer une seule matrice des taux de transitions comportant les probabilités de transition à une position donnée d'un état à l'autre pour toutes les séquences du tableau.

   réponse: (2pt)

	1	2	3	4	5
1		2/4	2/4		
2			2/2		
3				3/4	1/4
4			1/3		2/3
5		1/1			

3. Appliquer l'algorithme GSP à l'ensemble des données du tableau en utilisant un support minimum s = 33 % pour déterminer toutes les séquences fréquentes.

   réponse: (4pt)

1-itemsets sup count (1pt)	~~<(2), (2)> 0~~	~~<(5), (2)> 1~~	**3-itemsets sup count** (1pt)
< (1) > 4	<(2), (3)> 2	~~<(5), (3)> 1~~	<(1), (2), (3)> 2
< (2) > 2	<(2), (4)> 2	~~<(5), (4)> 1~~	<(1), (2), (4)> 2
< (3) > 4	~~<(2), (5)> 0~~	~~<(5), (5)> 0~~	<(1), (3), (4)> 3
< (4) > 4	~~<(3), (1)> 0~~	~~<(1, 2)> 0~~	<(1), (3), (5)> 2
< (5) > 4	~~<(3), (2)> 0~~	~~<(1, 3)> 0~~	~~<(1), (4), (5)> 1~~
2-itemsets sup count (1pt)	~~<(3), (3)> 1~~	~~<(1, 4)> 0~~	<(2), (3), (4)> 2
~~<(1), (1)> 0~~	<(3), (4)> 3	~~<(1, 5)> 1~~	~~<(2), (3), (5)>~~ (élagué)
<(1), (2)> 2	<(3), (5)> 2	~~<(2, 3)> 0~~	~~<(2), (4), (5)~~> (élagué)
<(1), (3)> 4	~~<(4), (1)> 0~~	~~<(2, 4)> 0~~	~~<(3), (4), (5)>~~ 1
	~~<(4), (2)> 0~~	~~<(2, 5)> 0~~	**4-itemsets sup count**
	~~<(4), (3)> 1~~	~~<(3, 4)> 0~~	

<(1), (4)> 3 <(1), (5)> 2 ~~<(2), (1)> 0~~	~~<(4), (4)> 0~~ <(4), (5)> 2 ~~<(5), (1)>~~ 0	~~<(3, 5)> 1~~ ~~<(4, 5)> 0~~	(1pt) <(1), (2), (3), (4)> 2

4. En déduire les motifs fréquents maximums.

   < (1), (2), (3), (4) >, < (1), (3), (5) >, < (4), (5) >(2pt)