Συστήματα ανακάλυψης γνώσεις από βάσεις Δεδομένων Ιανουάριος 2022

ΜΕΡΟΣ Α

1. Θεωρήστε το παρακάτω σύνολο δεδομένων εκπαίδευσης. Έχουμε 9 εγγραφές με το ύψος και το φύλο του κάθε ατόμου, οπού το φύλο είναι class variable.

Ύψος	161	164	169	175	176	179	181	184	185
Φύλο	F	F	M	M	F	F	M	M	F

   Το GINI INDEX του συνόλου των δεδομένων είναι:
   (α)1-(4/9)^2-(5/9)^2
   (β)(4/9)^2+(5/9)^2
   (γ)1-[(4/9)^2 – (5/9)^2]
2. Θεωρήστε το παρακάτω σύνολο δεδομένων εκπαίδευσης. Έχουμε 9 εγγραφές με το ύψος και το φύλο του κάθε ατόμου, οπού το φύλο είναι class variable.

Ύψος	161	164	169	175	176	179	181	184	185
Φύλο	F	F	M	M	F	F	M	M	F

   Αν επιλέξουμε να διασπάσουμε με βάση την τιμή 165, τότε
   (α)gini(<165)=1-(0/2)^2
   (β)gini(<165)=1-(2/9)^2
   (γ)gini(>165)=1-[(4/7)^2-(3/7)^2]
   (δ)gini(>165)=1-(4/7)^2-(3/7)^2
3. Στα προηγούμενα δεδομένα, αν τελικά επιλέξουμε το 165 ως ρίζα του δέντρου απόφασης και αποφασίσουμε ότι το δέντρο μας θα έχει μόνο έναν κόμβο , τότε τα δεδομένα εκπαίδευσης που το δέντρο κατηγοριοποιεί σωστά :
   (a)6
   (b)9
   (c)2
   (d)7
4. Υποθέστε ότι υπάρχουν 50 στιγμιότυπα της κλάσης P(positive) και 150 της κλάσης N(negative) σε ένα σύνολο δεδομένων ελέγχου 200 στιγμιότυπων. Έχουμε ένα κατηγοριοποίητη που κατηγοριοποιεί 40 στιγμιότυπα ως P από τα οποία την πραγματικότητα τα 30 ανήκουν στην κλάση P. Ποιες είναι οι τιμές του precision και recall;
   (α)30/40 και 30/50
   (β)30/50 και 30/40
   (γ)30/200 και 40/200
   (δ)50/150 και 30/150 (ίσως είναι 50/50 το πρώτο)
5. Υποθέστε ότι υπάρχουν 50 στιγμιότυπα της κλάσης P(positive) και 150 της κλάσης N(negative) σε ένα σύνολο δεδομένων ελέγχου 200 στιγμιότυπων. Έχουμε ένα κατηγοριοποίητη που κατηγοριοποιεί 40 στιγμιότυπα ως P από τα οποία την πραγματικότητα τα 30 ανήκουν στην κλάση P. Ποιο είναι το accuracy;
   (a)170/200
   (b)30/50
   (γ)40/50
   (d)80/100
6. Έστω τα σημεία α ως i στον μονοδιάστατο χώρο. Κάθε σημείο είναι Άσπρο(Α) ή Μαύρο(Μ)

coord	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
class		A	A		M		A		M						A		M			M
name		a	b		c		d		e						g		h			i

Αν εφαρμόσετε τον αλγόριθμο απομάκρυνσης θορύβου ENN με κ=3 στο dataset αυτό, τότε το ES(Edited Set) θα είναι το :

(a)ES={a,b,e,h,i}

(b)ES={a,b,r,g,i}

(c)ES={a,b,e,g,h}

(d)ES={a,b,d,h,i}

7. Έστω τα σημεία α ως i στον μονοδιάστατο χώρο. Κάθε σημείο είναι Άσπρο(Α) ή Μαύρο(Μ)

coord	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
class		A	A		M		A		M						A		M			M
name		a	b		c		d		e						g		h			i

Αν εφαρμόσετε τον αλγόριθμο IB2 στο dataset αυτό, τότε το CS(Edited Set) θα είναι το :

(a)CS={a,c,d,e,g,h}

(b)CS={a,c,d,e,g,i}

(c)CS={a,c,d,e,f,h}

(d)CS={a,b,d,e,g,h}

8. Έστω τα σημεία α ως i στον μονοδιάστατο χώρο. Σας δίνονται τα πέντε πρώτα βήματα της ιεραρχικής συσταδοποιησης με μέτρο απόστασης των συστάδων τη μέθοδο MAX distance(complete linkage). Ποιο είναι το βήμα 6; (Σε περίπτωση ισοπαλιών , επιλέξτε την επιλογή αριστερά)

coord	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
class		a	b		c		d		e		f				g		h			i

Στην εκφώνηση έδινε τα πρώτα βήματα και ζητούσε να επιλέξουμε το 6^{ο}

9. Έστω τα σημεία a ως i στον μονοδιάστατο χώρο. Εφαρμοστέ DBSCAN με minpoints=3 και epsilon=2

coord	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
class		a	b		c		d		e		f				g		h			i

Ποια είναι τα CORE POINTS;

(a)CORE={b,c,d,e}

(b)CORE={b,c,d,e,g}

(c)CORE={b,c,d,e,g,h}

(d)CORE={a,b,c,d,e}

10. Έστω τα σημεία a ως i στον μονοδιάστατο χώρο. Εφαρμοστέ DBSCAN με minpoints=3 και epsilon=2

coord	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
class		a	b		c		d		e		f				g		h			i

Ποια είναι τα NOISE POINTS;

(a)NOISE={g,h,i}

(b)NOISE={d,g,h,i}

(c)NOISE={g,i}

(d)NOISE={i}

11. Έστω τα σημεία a ως i στον μονοδιάστατο χώρο. Εφαρμοστέ DBSCAN με minpoints=3 και epsilon=2

coord	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
class		a	b		c		d		e		f				g		h			i

Ποια είναι τα BORDER POINTS;

(a)BORDER={a,f}

(b)BORDER={c,f}

(c)BORDER={a,g}

(d)BORDER={a,c,f,h}

ΜΕΡΟΣ Β [60 μονάδες]

1) Θέμα 1ο

Για τις παρακάτω δοσοληψίες και minsup=60% και mincof=80%

Δοσοληψία	Αντικείμενα
1	A, B, C, D, E, F
2	B, C, D, E, F, G
3	A, D, E, H
4	A, D, F, I, J
5	B, D, E, K

a) Κυκλώστε τα συχνά στοιχειοσυνολα
{A, B} {A, D} {A, E} {A, F} {B, D} {B, E} {B, F} {D, E} {D, F} {E, F} {A, B, D} {A, D, F} {B, D, E} {B, D, F} {D, E, F}

b) Γράψτε μόνο τους κανόνες(ή κανόνα) που πληρούν τους περιορισμούς και περιέχουν το D στο αριστερό μέρος .

c) Υπολογίστε και το interest/lift για τον/τους κανόνα/νες του Β.

2) ΘΕΜΑ 2^{Ο}

	DOC id	Λέξεις Εγγράφου	Μαγειρική
	1	φούρνος, τηγάνι, λάδι	Ναι
Σύνολο Εκπαίδευσης	2	φούρνος, φούρνος ,γάλα, ζάχαρη	Ναι
	3	πλυντήριο, ψυγείο, στεγνωτήριο	Όχι
	4	φούρνος, ψυγείο ,ψυγείο, πλυντήριο, στεγνωτήριο	Όχι
Σύνολο Ελέγχου	5	φούρνος, φούρνος, φούρνος	

a) Η κατηγορία του 5^{ο} εγγράφου με χρήση του κατηγοριοποιητή Naïve Bayes. Γράψτε τους υπολογισμούς σας.
b) Η κατηγορία του 5^{ο} εγγράφου με χρήση του κατηγοριοποιητή Binary Multinomial Naïve Bayes. Γράψτε τους υπολογισμούς σας.

2) ΘΕΜΑ 3^{Ο}

a) Θεωρήστε το παρακάτω web log:

#	IP Address	TIME	URL		Agent
1	IP2	9/Nov/05:03:05:06	GET A.HTML	--	Agent2
2	IP1	9/Nov/05:03:05:26	GET A.HTML	--	Agent1
3	IP2	9/Nov/05:03:06:06	GET X.HTML	A.HTML	Agent2
4	IP2	9/Nov/05:03:06:39	GET B.HTML	A.HTML	Agent2
5	IP1	9/Nov/05:03:07:03	GET C.HTML	A.HTML	Agent1
6	IP1	9/Nov/05:03:07:20	GET D.HTML	C.HTML	Agent1
7	IP1	9/Nov/05:03:08:40	GET E.HTML	C.HTML	Agent1
8	IP1	9/Nov/05:03:29:06	GET W.HTML	A.HTML	Agent1
9	IP2	9/Nov/05:04:10:06	GET Z.HTML	--	Agent2
10	IP2	9/Nov/05:04:15:46	GET O.HTML	Z.HTML	Agent2

Διαιρέστε το log sessions ανά χρήση, αφού εντοπίσετε και τους διαφορετικούς χρήστες, χρησιμοποιώντας time out 5 λεπτών για παραμονή στην ιδιά σελίδα. Η απάντηση θα είναι της μορφής Χρήστης 1- Session 1:1,2,3 (όπου οι αριθμοί αντιστοιχούν στις γραμμές του log), Χρήστης 1-Session 2: 4,5,6

b) Επιλέξτε τα συμπληρωμένα μονοπάτια που αντιστοιχούν σε sessions του A
   i) A->C->D->C->A->E
   ii) A->C->A->E
   iii) A->C->D->C->A->E->W
   iv) A->C->A->E->W
   v) A->X
   vi) A->X->B->C->D->E->W

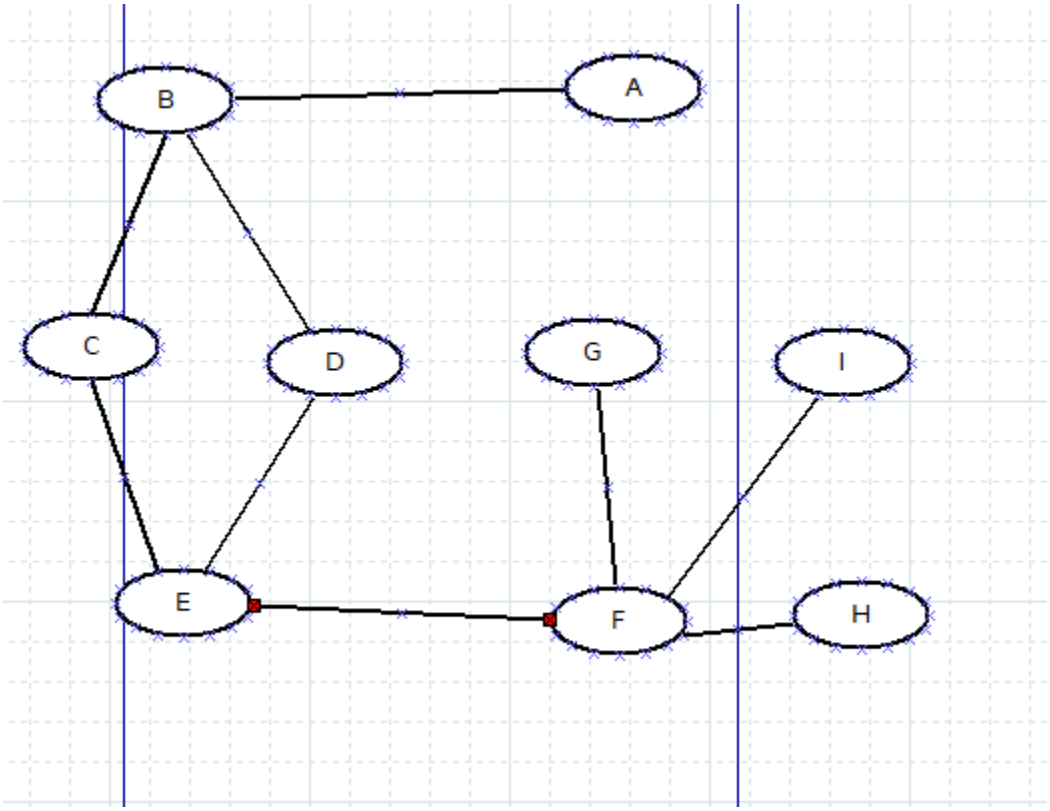
vii) Z->D
viii) A->X->A->B
ix) A->W
x) C->D
xi) A->X->B
xii) A->X->B->D->E->W

3) ΘΕΜΑ 4ο
   a) Θεωρήστε τον γράφο που δίνεται από τον παρακάτω πίνακα γειτνίασης

	P1	P2	P3	P4
P1	0	1	1	0
P2	1	0	1	1
P3	0	1	0	0
P4	1	0	1	1

   Γράψτε με διατύπωση πινάκων την 1η επανάληψη για τα power iterations του PageRank (δηλαδή την διατύπωση με χρήση πινάκων M και r , χωρίς να κάνετε τον υπολογισμό)
   b) Γράψτε με διατύπωση πινάκων την 1η επανάληψη για τα power iterations του PageRank με β=0,8 (παρόμοια με το Α)

4) ΘΕΜΑ 5ο
   a) Για το παρακάτω γράφο



   Υπολογίστε το edge betweenness για την ακμή που έχει την μεγαλύτερη τιμή

b) Ταξινομήστε τις ακμές σε φθίνουσα σειρά με βάση το edge betweenness. Σε περίπτωση ισοβάθμιας προηγείται η ακμή με το μικρότερο λεξικογραφικό άκρο. Δεν απαιτείται ο υπολογισμός όλων των τιμών αν δεν θέλετε.
c) Δώστε τα ιεραρχικά clusters που θα προκύψουν με την εφαρμογή του αλγορίθμου Girvan-Newman και αν υποθέσουμε ότι εκμεταλλευόμαστε τα αποτελέσματα του Β, χωρίς επαναυπολογισμό του edge betweenness σε κάθε βήμα. Για την αναπαράσταση χρησιμοποιείστε δενδρόγραμμα.