ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟΥΠΟΛΗ – ΡΙΟ 26504 ΠΑΤΡΑ

ΕΦΟ-03 / ΔΙΑΧΕΙΡΙΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ
Εξεταστική Περίοδος Φεβρουαρίου Ακαδημαϊκού Έτους 2018-2019

Ονοματεπώνυμο:……

Α. Μ. : ……

Θέματα **(30 ερωτήσεις)**

1. Έστω ένα σύνολο δεδομένων με 28 μεταβλητές που όλες λαμβάνουν αριθμητικές τιμές. Αν εφαρμοστεί ο αλγόριθμος Ανάλυσης Κύρων Συνιστωσών (Principal Component Analysis – PCA) με χρήση του πίνακα συνδιακύμανσης σε αυτό το σύνολο δεδομένων, πόσα ιδιοδιανύσματα (eigenvectors) και πόσες ιδιοτιμές (eigenvalues) θα προκύψουν;

Α	**Β**	**Γ**	**Δ**
28 ιδιοδιανύσματα και 14 ιδιοτιμές	28 ιδιοδιανύσματα και 28 ιδιοτιμές	14 ιδιοδιανύσματα και 14 ιδιοτιμές	Καμία από τις απαντήσεις Α, Β, Γ

2. Δίνεται ο παρακάτω πίνακας δεδομένων δύο μεταβλητών (M1 και M2) και 3 παρατηρήσεων:

M1	**M2**
1	green
5	blue
6	red

Αν εφαρμοστεί στα παραπάνω δεδομένα ο αλγόριθμος Ανάλυσης Κυρίων Συνιστωσών (PCA) με τη χρήση πίνακα συνδιακύμανσης (covariance matrix), τα *ιδιοδιανύσματα* που θα προκύψουν είναι:

Α	**Β**	**Γ**	**Δ**
$\begin{pmatrix}-0.082\\1\end{pmatrix}\begin{pmatrix}12.106\\1\end{pmatrix}$	$\begin{pmatrix}-9.81\\-1.45\end{pmatrix}\begin{pmatrix}1.43\\1\end{pmatrix}$	$\begin{pmatrix}-0.8\\-1.4\end{pmatrix}\begin{pmatrix}-0.43\\-2.5\end{pmatrix}$	Κανένα από τα Α, Β, Γ

3. Θέλουμε να συλλέξουμε αναρτήσεις χρηστών στο Twitter και να εξετάσουμε ποιες από τις αναρτήσεις αυτές εκφράζονται αρνητικά, θετικά ή ουδέτερα για την Ελένη Μενεγάκη. Οι αλγόριθμοι εξόρυξης δεδομένων που μπορούν να χρησιμοποιηθούν για να λυθεί αυτό το πρόβλημα είναι: *(επιλέξτε παραπάνω από μία)*

Α	Δέντρα απόφασης
Β	K-means
Γ	Apriori
Δ	Naïve Bayes
Ε	K-Nearest Neighbors

4. Σε ένα δυαδικό πρόβλημα κατηγοριοποίησης με δέντρα απόφασης, όπου η ετικέτα κατηγορίας (class attribute) μπορεί να λάβει μόνο δυο διακριτές τιμές (π.χ. ”yes”/ “no”, ”tall”/”short”, “cheat” / ”no-cheat” κλπ), η μέγιστη τιμή της εντροπίας ενός κόμβου είναι:

Α	**Β**	**Γ**	**Δ**	**Ε**
0	0.3	0.5	0.75	1

5. Δίνεται ο παρακάτω πίνακας, που αναπαριστά έναν κόμβο σε δέντρο απόφασης, με το γνώρισμα «Ευτυχισμένος» να είναι η ετικέτα κατηγορίας (class attribute).

Ηλικία	**Εισόδημα**	**Ευτυχισμένος**
25	Υψηλό	Όχι
36	Υψηλό	Ναι
40	Χαμηλό	Όχι
27	Μεσαίο	Όχι

Ο συντελεστής Gini του πάραπάνω κόμβου είναι:

Α	**Β**	**Γ**	**Δ**	**Ε**
0.275	0.456	0.375	0.5	1

6. Δίνεται ο παρακάτω πίνακας, που αναπαριστά έναν κόμβο σε δέντρο απόφασης, με το γνώρισμα «Εισόδημα» να είναι η ετικέτα κατηγορίας (class attribute).

Ηλικία	**Οικογενειακή κατάσταση**	**Εισόδημα**
25	Άγαμος	Υψηλό
36	Διαζευγμένος	Μεσαίο
40	Παντεμένος	Μεσαίο
27	Παντρεμένος	Χαμηλό
28	Άγαμος	Χαμηλό

Η εντροπία (Entropy) του παραπάνω κόμβου είναι:

Α	**Β**	**Γ**	**Δ**	**Ε**
0.672	0.762	0.124	1.521	1.023

7. Δίνεται ο παρακάτω πίνακας δεδομένων εκπαίδευσης, με το γνώρισμα C να είναι η ετικέτα κατηγορίας (class attribute). Επιθυμούμε να χτίσουμε ένα δέντρο απόφασης.

A	**B**	**C**
A1	B1	C1
A1	B2	C2
A2	B1	C2
A3	B1	C1

Με ποιο γνώρισμα πρέπει να γίνει η πρώτη διάσπαση των παραπάνω δεδομένων (κόμβου), αν γίνει χρήση του συντελεστή Gini και του Gini gain?

Α	**Β**	**Γ**	**Δ**
Γνώρισμα Α	Γνώρισμα Β	Γνώρισμα C	Οποιοδήποτε από τα γνωρίσματα Α ή Β

8. Στα δέντρα απόφασης, με τον όρο «Υπερπροσαρμογή μοντέλου» (Overfitting) περιγράφεται η κατάσταση κατα την οποία:

Α	Τα σφάλματα εκπαίδευσης είναι μικρά και τα σφάλματα γενίκευσης είναι μικρά
Β	Τα σφάλματα εκπαίδευσης είναι μεγάλα και τα σφάλματα γενίκευσης είναι μεγάλα
Γ	Τα σφάλματα εκπαίδευσης είμαι μεγάλα ενώ τα σφάλματα γενίκευσης είναι μικρά
Δ	Τα σφάλματα εκπαίδευσης είναι μικρά ενώ τα σφάλματα γενίκευσης είναι μεγάλα.
Ε	Δεν υπάρχουν ούτε σφάλματα εκπαίδευσης ούτε σφάλματα γενίκευσης

9. Δίνεται ο παρακάτω πίνακας δεδομένων με το γνώρισμα “Cheat” να είναι η ετικέτα κατηγορίας (class attribute).

Debt	Marital status	Cheat
High	Divorced	No
Low	Married	Yes
High	Married	Yes
Low	Single	No

Αν το παραπάνω σύνολο δεδομένων είναι το σύνολο εκπαίδευσης, σε ποια κατηγορία πρέπει να μπει το άγνωστο δεδομένο με τιμές *(Low, Divorced)* στα γνωρίσματα Debt και Marital status αντίστοιχα, αν εφαρμοστεί ο αλγόριθμος K-Nearest Neighbors (K-NN) με K=3 και μέτρο απόστασης την απόσταση Overlap (Hamming) ?

Α	**Β**	**Γ**
Yes	No	Οποιοδήποτε από τα Yes, No.

10. Κατά την εκτέλεση του αλγορίθμου K-Nearest Neighbors (K-NN), που κάνει χρήση της Ευκλείδειας απόστασης, πρέπει να γίνεται κανονικοποίηση (scaling, normalization) όλων των συνεχών αριθμητικών γνωρισμάτων των δεδομένων με τη χρήση των μεθόδων min-max normalization ή z-score επειδή:

Α	Για να βρίσκεται η Ευκλείδεια απόσταση μεταξύ των δεδομένων στο διάστημα (0,1).
Β	Για να βρίσκεται η Ευκλείδεια απόσταση μεταξύ των δεδομένων στο διάστημα (-3,3).
Γ	Για να μην προκύπτουν πολύ μικρές τιμές της Ευκλείδειας απόστασης κατά τη διάρκεια υπολογισμών.
Δ	Για να μην επηρεάζουν δυσανάλογα την Ευκλείδεια απόσταση γνωρίσματα με μεγάλο εύρος τιμών.
Ε	Δεν χρειάζεται ποτέ να γίνεται τέτοια κανονικοποίηση των συνεχών αριθμητικών γνωρισμάτων στην περίπτωση του αλγορίθμου K-NN.

11. Δίνεται ο παρακάτω πίνακας δεδομένων εκπαίδευσης, με το γνώρισμα C να είναι η ετικέτα κατηγορίας (class attribute).

A	B	C
A1	B1	C1
A1	B2	C2
A2	B1	C2
A3	B1	C1
A2	B2	C1

A2	B1	C2
A3	B3	C3

Σε ποιά κατηγορία θα πρέπει να μπει το άγνωστο δεδομένο με τιμές *(A2, B2)* στις μεταβλητές Α, Β αντίστοιχα, αν η κατηγοριοποίηση γίνει με τον αλγόριθμο Naïve Bayes?

Α	**Β**	**Γ**	**Δ**
C1	C2	C3	Οποιαδήποτε από τις C1, C2, C3

12. Στη βιβλιοθήκη scikit-learn της Python, διατίθενται τρεις συναρτήσεις για τη δημιουργία κατηγοριοποιητή Naïve Bayes: *GaussianNB(), MultinomialNB() και BernoulliNB()*. Αναφέρετε συνοπτικά σε ποιες περιπτώσεις χρησιμοποιείται κάθε μία από τις παραπάνω συναρτήσεις.

GaussianNB() :

MultinomialNB() :

BernoulliNB() :

13. Δίνεται ο παρακάτω πίνακας δεδομένων για πελάτες τράπεζας, με το το γνώρισμα Risk, να είναι ετικέτα κατηγορίας (class attribute)

Unemployed	**Owns House**	**Age**	**Risk**
Yes	Yes	Young	High
Yes	No	Old	High
Yes	No	Middle	Low
No	No	Middle	Medium
No	Yes	Old	High
Yes	Yes	Old	Medium
No	Yes	Young	Low

Η απόσταση της τιμής 'Young' από την τιμή 'Old' του γνωρίσματος Age - δηλαδή η τιμή *d("Young", "Old")* - αν χρησιμοποιηθεί η μετρική Modified Value Difference Metric (MVDM), είναι:

Α	**Β**	**Γ**	**Δ**	**Ε**
0.5	0.857	0.833	1.375	2.543

14. Θέλουμε να υπολογίσουμε την απόσταση μεταξύ των εξής δύο διανυσμάτων:

I. (33, 1.234, Audi, TRUE, 0.567, Tyrian Purple)
II. (19. 0.873, Fiat, FALSE, 0.567, Navy Blue)

Ποιά από τις παρακάτω μετρικές αποστάσεων μπορούν να χρησιμοποιηθούν?

A	Μπορεί να χρησιμοποιηθεί η Ευκλείδια απόσταση
B	Μπορεί να χρησιμοποιηθεί η απόσταση Hamming
C	Μπορεί να χρησιμοποιηθεί η Modified Value Difference Metric (MVDM)

D	Κανένα από τα παραπάνω

15. Δίνονται τα παρακάτω μοντέλα παλινδρόμησης, όπου y η εξαρτημένη μεταβλητή, x_i η ανεξάρτητη και b_i οι συντελεστές του μοντέλου. Ποια από τα παρακάτω μοντέλα παλινδρόμησης είναι γραμμικά και ποια μη-γραμμικά; Σημειώστε δίπλα σε κάθε μοντέλο Γ αν είναι γραμμικό και ΜΓ αν είναι μη-γραμμικό:

Μοντέλο	**Γραμμικό (Γ) ή μη-γραμμικό (ΜΓ)**
$y = b_1 x_1^3 + b_2 x_2^3 + b_0$	
$\ln(y) = b_1 x_1^2 + b_2 \frac{1}{x_2^4} + b_0$	
$y = b_1 x_1 + b_2 e^{-b_3 x_2} + b_0$	
$\ln(y) = b_1 \ln(x_1) + b_2 \ln(x_2) + b_0$	

16. Στη μέθοδο της σταδιακής καθόδου (gradient descent) η ενημέρωση των συντελεστών θ_j δίνεται από τον τύπο:

$$\theta_j = \theta_j - a \frac{\partial J(\theta)}{\partial \theta_j}$$

όπου α η παράμετρος μάθησης.

Εξηγείστε γιατί η ποσότητα $a \frac{\partial J(\theta)}{\partial \theta_j}$ *πρέπει να αφαιρεθεί από την τρέχουσα τιμή θj* και όχι π.χ. να προστεθεί σε αυτήν για την ενημέρωση των συντελεστών στη μέθοδο της σταδιακής καθόδου.

17. Αναφέρετε τρεις διαφορές της μεθόδου σταδιακής καθόδου (Gradient Descent) από τη μέθοδο των ελαχίστων τετραγώνων (Ordinary Least Squares – OLS) για την εκτίμηση των συντελεστών μοντέλων παλινδρόμησης.

18. Δίνονται οι παρακάτω 4 δοσοληψίες μιας βάσης δεδομένων καλαθιού αγοράς

Tid	Items
1	Γάλα, Πάνες, Μπύρα
2	Μπύρα, Οδοντόκρεμα, Απορρυπαντικό, Coca-Cola
3	Coca-Cola, Γαριδάκια, Μπύρα
4	Γαριδάκια, Μπύρα, Χαρτί υγείας, Πορτοκαλάδα, Coca-Cola

Η υποστήριξη (support of itemset, s) του 2-στοιχειοσυνόλου *{Cola-Cola, Γαριδάκια}* είναι:

Α	Β	Γ	Δ	Ε
0.75	1	0.5	0.25	0

19. Δίνονται οι παρακάτω 4 δοσοληψίες μιας βάσης δεδομένων καλαθιού αγοράς

Tid	Items
1	Γάλα, Πάνες, Μπύρα, Coca-Cola
2	Μπύρα, Οδοντόκρεμα, Πορτοκαλάδα, Απορρυπαντικό, Coca-Cola
3	Coca-Cola, Γαριδάκια, Μπύρα
4	Γαριδάκια, Μπύρα, Χαρτί υγείας, Πορτοκαλάδα, Coca-Cola

Υπολογίστε την υποστήριξη s (support) και την εμπιστοσύνη c (confidence) του κανόνα συσχέτισης

{Μπύρα}* → *{Coca-Cola, Πορτοκαλάδα}

βάσει της παραπάνω βάσης δεδομένων καλαθιού αγοράς

20. Δίνονται οι παρακάτω 3 δοσοληψίες μιας βάσης δεδομένων καλαθιού αγοράς που περιέχουν τα στοιχεία I1, I2, I3 και I4

Tid	Items
1	I1, I2, I3
2	I2, I3
3	I1, I2, I4

Αν εφαρμοστεί ο αλγόριθμος Apriori στην παραπάνω βάση δεδομένων δοσοληψιών, με ελάχιστη μέτρηση υποστήριξης 2 (support count σ = 2), ποια συχνά στοιχειοσύνολα θα προκύψουν?

Α	**Β**	**Γ**	**Δ**	**Ε**
{I1} {I2} {I3} {I1, I2}, {I2,I3}	{I1} {I4} {I1, I2} {I2,I4}	{I2} {I3} {I4} {I2,I3} {I1, I2, I3}	{I1} {I3} {I1, I4} {I1, I2,I3}	{I2} {I3} {I2,I3}

21. Θεωρήστε ότι εφαρμόζεται ο αλγόριθμος Apriori σε μία βάση δεδομένων δοσοληψιών που περιέχει τα στοιχεία 1,2,3,4 και 5. Κατά την εφαρμογή του αλγορίθμου Apriori προέκυψαν τα εξής **συχνά 3-στοιχειοσύνολα:** {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5} .
Βρείτε ποια από τα παρακάτω υποψήφια 4-στοιχειοσύνολα επιβιώνουν από το βήμα περικοπής (κλαδέματος) υποψηφίων του Apriori αλγορίθμου όταν θα υπολογιστούν τα συχνά 4-στροιχειοσύνολα:

Α	**Β**	**Γ**	**Δ**	**Ε**
{1, 2, 3, 4} και {1, 2, 4, 5}	Μόνο το {1, 2, 3, 4}	{1, 2, 3, 4} και {1, 2, 3, 5}	{1, 2, 4, 5}	Κανένα από τα προηγούμενα

22. Έστω ένα μικρό σύνολο δεδομένων, το οποίο περιέχει 4 δοσοληψίες, οι οποίες περιέχουν στοιχεία από το 1 έως το 5, όπως φαίνεται παρακάτω:

Tid	**Items**
1	1,3,4
2	2,3,5
3	1,2,3,5
4	2,5

Θεωρώντας ως ελάχιστη μέτρηση υποστήριξης 2 (support count σ=2), τα *μέγιστα συχνά στοιχειοσύνολα (Maximal frequent itemsets)* είναι:

Α	**Β**	**Γ**	**Δ**	**Ε**
{1,2} {2,3,4}	{2,4} {1,2,3}	{1,3} {2,4} {1,3,5}	{1,3}, {2,3,5}	{1,2} {4,5} {3,4,5}

23. Σε μία βάση δεδομένων δοσοληψιών, η εμπιστοσύνη του κανόνα συσχέτισης *{X} → {Y,Z}* βρέθηκε ότι είναι 0.85. Η τιμή 0.85 σημαίνει ότι:

Α	Το 85% των δοσοληψιών της βάσης δεδομένων περιέχουν τα X, Y και Z.
Β	Το 85% των δοσοληψιών της βάση δεδομένων περιέχουν τουλάχιστον ένα από τα X, Y και Z.
Γ	Το 85% των δοσοληψιών της βάσης δεδομένων που περιέχουν τα Y και Z περιέχουν επίσης και το X.
Δ	Το 85% των δοσοληψιών της βάσης δεδομένων που περιέχουν το X, περιέχουν, επίσης, και τα Y και Z.
Ε	Τίποτα από τα παραπάνω

24. Δίνονται 5 σημεία P1 έως και P5 στον δισδιάστατο χώρο, των οποίων οι συντεταγμένες (x,y) έχουν ως εξής:

	x	*y*
P1	0.0	3.0

P2	2.0	1.0
P3	2.0	2.0
P4	4.5	0.0
P5	5.0	2.0

Εφαρμόστε συσσωρευτική ιεραρχική συσταδοποίηση (agglomerate hierarchical clustering) χρησιμοποιώντας ως μέτρο απόστασης συστάδων την μέθοδο MIN (single linkage – απλού συνδέσμου) και απεικονίστε στον παρακάτω χώρο το δεντρόγραμμα των συστάδων που προκύπτει. Στον κάθετο άξονα του δεντρογράμματος θα πρέπει να φαίνονται οι αποστάσεις των συστάδων που δημιουργούνται.

25. Στους κανόνες συσχέτισης, η έννοια της ανύψωσης (Lift) εκφράζει:

Α	Πόσο αξιόπιστος είναι ένας κανόνας συσχέτισης
Β	Τί κάλυψη έχει ένας κανόνας συσχέτισης
Γ	Την εμπιστοσύνη ενός κανόνα συσχέτισης
Δ	Πόσο ενδιαφέρον είναι ένας κανόνας συσχέτισης
Ε	Πόσο λογικός είναι ένας κανόνας συσχέτισης

26. Έστω ένας κανόνας συσχέτισης $X \rightarrow Y$, με τιμή ανύψωσης (Lift) ίση με 2.34 . Αυτή η τιμή ανύψωσης σημαίνει ότι:

Α	Τα στοιχειοσύνολα Χ και Υ συνεμφανίζονται στις δοσοληψίες λιγότερες φορές απ'ότι αναμενόταν (αρνητική συσχέτιση).
Β	Τα στοιχειοσύνολα Χ και Υ συνεμφανίζονται στις δοσοληψίες ακριβώς όσες φορές αναμενόταν (καμία συσχέτιση).
Γ	Τα στοιχειοσύνολα Χ και Υ συνεμφανίζονται στις δοσοληψίες περισσότερες φορές απ'ότι αναμενόταν (θετική συσχέτιση).

Δ	Δεν μας λέει τίποτε για τον τρόπο συσχέτισης των στοιχειοσυνόλων Χ και Υ.

27. Δίνεται ο παρακάτω πίνακας μέτρησης συχνότητας εμφάνισης (contingency table) στοιχειοσυνόλων σε μία βάση δεδομένων δοσοληψιών

	$Chair, Shampoo$	$\overline{Chair, Shampoo}$	
$Coffee$	16	15	31
$\overline{Coffee}$	51	12	63
	67	27	94

Η ανύψωση (lift) του κανόνα συσχέτισης **{Coffee}** → **{Chair, Shampoo}** είναι:

Α	**Β**	**Γ**	**Δ**	**Ε**
0.724	0.998	1.034	2.782	0.432

28. Δίνονται τα παρακάτω δεδομένα μίας βάσης δεδομένων πελατών μιας τράπεζας

Κωδικός ID	**Income**	**Number children**	**Debt**
1	23000	2	2000
2	12000	1	5000
3	54000	0	10000
4	52000	3	8000

Στα παραπάνω δεδομένα εφαρμόζεται ο αλγόριθμος K-means με στόχο να χωριστούν τα δεδομένα σε δύο (2) συστάδες (K=2) και με αρχικά επίκεντρα συστάδων τα δεδομένα με κωδικό (Κωδικός, ID) 1 και 2 του πίνακα παραπάνω. Αναφέρετε ποια θα είναι τα νέα επίκεντρα των δύο συστάδων που θα προκύψουν **μετά την πρώτη επανάληψη** του αλγορίθμου K-means.

29. Τί κάνει ο παρακάτω κώδικας R?

```
n <- function(x) { return( (x-min(x) ) / (max(x)-min(x)) ) }

mydata[,"Income"] <- n( mydata$Income )
```

```
mydata[,"Age"] <- n( mydata$Age )
```

Α	Κάνει κανονικοποίηση (scaling, normalization) με τη χρήση της μεθόδου min-max των μεταβλητών Income και Age των δεδομένων mydata
Β	Κάνει κανονικοποίηση (scaling, normalization) με τη χρήση της μεθόδου z-score των μεταβλητών Income και Age των δεδομένων mydata
Γ	Κάνει κανονικοποίηση (scaling, normalization) με τρόπο ώστε οι τιμές των γνωρισμάτων Income και Age των δεδομένων mydata να απεικονιστούν στο διάστημα [-3,3]
Δ	Τίποτα από τα παραπάνω

30. Στη γλώσσα προγραμματισμού R, η εκτέλεση της εντολής

```
someData <- c(9, -21, 33, -4, 0.4)
```

δημιουργεί:

Α	Ένα νέο DataFrame που έχει 5 μεταβλητές και μία γραμμή δεδομένων με μία τιμή για κάθε μια από τις 5 μεταβλητές.
Β	Έναν νέο πίνακα (matrix) διαστάσεων 3 γραμμών και 2 στηλών (3x2).
Γ	Μία νέα λίστα (list) με 5 στοιχεία.
Δ	Ένα νέο διάνυσμα (vector) με 5 στοιχεία.
Ε	Τίποτα από τα παραπάνω

Καλή Επιτυχία!
04/02/2019 AD

BONUS Ερωτήσεις

Διαλέξτε 2 από τις παρακάτω 4 ερωτήσεις της αρεσκείας σας και απαντήστε τις:

1) Αποδείξτε ότι για κάθε αριθμό a > 0 ισχύει ότι

$$log_2(a) = \frac{\ln(a)}{\ln(2)}$$

όπου ln(α) ο λογάριθμος με βάση e.

2) Θα προτιμούσατε να πολεμήσετε 1 πάπια σε μέγεθος αλόγου ή 200 άλογα σε μέγεθος πάπιας? Αναπτύξτε.
3) Αποδείξτε ότι η πιθανότητα να πιάσεις τζακ-ποτ στο Τζόκερ είναι ίση με την πιθανότητα να κληρωθούν οι ίδιοι αριθμοί σε δύο διαδοχικές κληρώσεις στο Τζόκερ.
4) Η αγγλική λέξη 'Fat' μοιάζει σαν κάποιος να δάγκωσε τη λέξη 'Eat'. Αναπτύξτε.