ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΑΣ

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης
Υπεύθυνοι Καθηγητές: *Μεγαλοοικονόμου Βασίλειος, Μακρής Χρήστος*

ΘΕΜΑ 1 (25%)

α. Περιγράψτε τα βασικά χαρακτηριστικά του Partitioning Around Medoids (PAM) αλγορίθμου ομαδοποίησης. Ο συγκεκριμένος αλγόριθμος υπολογίζει τοπικά ή ολικά ελάχιστα; Εξηγείστε τις διαφορές του PAM από τον k-means.

β. Ορίστε το πρόβλημα της εξαγωγής κανόνων συσχέτισης όταν έχουμε να χειριστούμε πολλαπλές τιμές ελάχιστης υποστήριξης και εξηγείστε πως μπορούμε να τροποποιήσουμε τον αλγόριθμο Apriori για να χειριστεί κανόνες συσχέτισης με πολλαπλές τιμές ελάχιστης υποστήριξης.

γ. Στο πρόβλημα της εξαγωγής γνώσης από τον παγκόσμιο ιστό θέλουμε να εξάγουμε πληροφορίας από τις προσπελάσεις χρηστών σε ένα δικτυακό τόπο ώστε να προβλέψουμε τις μελλοντικές προσπελάσεις τους. Εξηγείστε τα βασικά αλγοριθμικά βήματα μίας τέτοιας προσέγγισης.

ΘΕΜΑ 2 (25%)

α. Χρησιμοποιείστε τον ακόλουθο πίνακα ομοιότητας για να πραγματοποιήσετε single και complete link hierarchical clustering. Επιδείξτε τα αποτελέσματά σας σχεδιάζοντας ένα δενδρόγραμμα. Το δενδρόγραμμα θα πρέπει να επιδεικνύει τη σειρά με την οποία τα σημεία συγχωνεύτηκαν.

	p1	p2	p3	p4	p5
p1	1.00	0.8	0.5	0.4	0.1
p2	0.8	1.00	0.55	0.42	0.35
p3	0.5	0.55	1.00	0.4	0.5
p4	0.4	0.42	0.4	1.00	0.8
p5	0.1	0.35	0.5	0.8	1.00

b. Περιγράψτε ένα αλγοριθμικό σχήμα για την παραγωγή κανόνων συσχέτισης που θα καταγράφει όχι μόνο την παρουσία αλλά και την απουσία στοιχειοσυνόλων. Εξηγείστε πως ο αλγόριθμος Apriori μπορεί να τροποποιηθεί για να χειριστεί τέτοιου είδους στοιχειοσύνολα.

c. Στο πρόβλημα της απόκρυψης κανόνων συσχέτισης θέλουμε να αφαιρέσουμε από τα συχνά συνολοστοιχεία ευαίσθητη πληροφορία (με τη μορφή συχνών συνολοστοιχείων). Πιο τυπικά αν F είναι τα συχνά συνολοστοιχεία θέλουμε να αποκρύψουμε ένα υποσύνολο των συνολοστοιχείων (ας το ονομάσουμε S) τα οποία θεωρούνται ευαίσθητα και αυτό μπορούμε να το επιτύχουμε με αφαίρεση στοιχείων από συναλλαγές ή προσθήκες συναλλαγών με σκοπό την ελαχιστοποίηση του αριθμού στην απόκρυψη. Περιγράψτε τις λεπτομέρειες μία τέτοιας προσέγγισης, και τα αλγοριθμικά προβλήματα που ανακύπτουν.

ΘΕΜΑ 3 (15%)

3.1) Ποια είναι τα βασικά βήματα προεπεξεργασίας των δεδομένων;
3.2) Για τη συμπλήρωση ελλιπών τιμών που υπάρχουν στα δεδομένα έχουν προταθεί διάφορες τεχνικές. Συμπληρώστε τις ελλιπείς τιμές με τις εξής μεθόδους:

i. με χρήση καθολικής μεταβλητής
ii. με τη μέση τιμή του χαρακτηριστικού
iii. με τη μέση τιμή του γνωρίσματος για τις πλειάδες της ίδιας κλάσης.

	Κλάση	Χαρακτηριστικό_1	Χαρακτηριστικό_2	Χαρακτηριστικό_3
Στοιχείο_1	1	56	16	12
Στοιχείο_2	0	64		
Στοιχείο_3	0	87	79	59
Στοιχείο_4	1	23	20	
Στοιχείο_5	1	44		24

ΘΕΜΑ 4 (20%)

4.1) Ποια είναι τα πλεονεκτήματα και μειονεκτήματα των Δένδρων Αποφάσεων;

4.2) Έστω η παρακάτω βάση δεδομένων με ποδήλατα, τα οποία αναπαριστώνται από 5 παραδείγματα εκπαίδευσης. Το ποδήλατο μπορεί να είναι αποδεκτό ή όχι (μεταβλητή κλάσης). Το αν ένα ποδήλατο είναι αποδεκτό ή όχι εξαρτάται απ' τα προηγούμενα 3 χαρακτηριστικά. Με βάση τα δεδομένα του πίνακα, να κατασκευαστεί ένα δένδρο απόφασης για την κατηγοριοποίηση των ποδηλάτων, στο οποίο οι υψηλότεροι κόμβοι θα αντιστοιχούν σε γνωρίσματα που αποδίδουν μεγαλύτερο κέρδος πληροφορίας. Περιγράψτε αναλυτικά τη διαδικασία κατασκευής του δένδρου απόφασης.

	Ανεξάρτητες μεταβλητές			**Εξαρτημένη μεταβλητή**
Ποδήλατο	**Αναρτήσεις**	**Ταχύτητες**	**Φως**	**Αποδεκτό**
1	Όχι	6	Ναι	Ναι
2	Ναι	6	Όχι	Όχι
3	Όχι	18	Όχι	Ναι
4	Όχι	6	Όχι	Όχι
5	Ναι	18	Όχι	Ναι

ΘΕΜΑ 5 (15%)

5.1) Ποια είναι τα βασικά είδη ερωτήσεων σε χωρικά δεδομένα;

5.2) Χρησιμοποιήστε τη μέθοδο Dynamic Time Warping (DTW) για τον υπολογισμό της απόστασης των χρονοσειρών X = {2,6,5,3} και Y = {2,3,5,2}. Εντοπίστε το Warping Path για τις χρονοσειρές X και Y.

5.3) Ποια είναι η διαφορά της μεθόδου DTW απ' τη μέθοδο Minimal Variance Matching (MVM);