ΚΕΦΑΛΑΙΟ 24

Κατηγοριοποίηση

Ασκήσεις - Ερωτήσεις

1. Έστω ότι στο παράδειγμα κατηγοριοποίησης στο Σχήμα 24.3 εφαρμόζεται στην είσοδο το διάνυσμα δεδομένων (0 1 ? 1 1 ?).
   α) Ποιες από τις λύσεις απορρίπτονται;
   β) Ποιες από τις λύσεις είναι συνεπείς;
   γ) Ποιες από τις λύσεις εξηγούν τα δεδομένα;

2. Στο ίδιο παράδειγμα (Σχήμα 24.3), ποια είναι η μορφή των διανυσμάτων δεδομένων για τα οποία:
   α) Η λύση S_2 είναι συνεπής.
   β) Η λύση S_2 εξηγεί τα δεδομένα.

   Σημείωση: Χρησιμοποιήστε το σύμβολο ? για να δηλώσετε άγνωστη τιμή για το δεδομένο και το σύμβολο * για να δηλώσετε ότι κάποιο στοιχείο μπορεί να πάρει οποιαδήποτε τιμή από τις {0, 1, ?}. Για παράδειγμα, το διάνυσμα (* (1 0) 1 ? 0 (0 ?)) δηλώνει ότι το πρώτο στοιχείο μπορεί να είναι οποιοδήποτε, το δεύτερο μπορεί να είναι 0 ή 1, το τρίτο οπωσδήποτε 1, το τέταρτο είναι άγνωστο, το πέμπτο 0 και το έκτο μπορεί να είναι 0 ή άγνωστο.

3. Πολλές φορές, τα προβλήματα επιλογής ενός ή περισσοτέρων αντικειμένων βάσει προτιμήσεων, μπορούν να αντιμετωπιστούν ως προβλήματα κατηγοριοποίησης. Συγκεκριμένα, κάθε διαθέσιμη επιλογή μπορεί να θεωρηθεί ως κατηγορία, ενώ οι προτιμήσεις αυτού που επιλέγει μπορούν να θεωρηθούν ως το διάνυσμα των δεδομένων. Με βάση την παρατήρηση αυτή, να αναπτυχθεί ένα σύστημα γνώσης "Μεσίτης" σε εργαλείο ή γλώσσα της προτίμησής σας, το οποίο να συμβουλεύει κάποιον υποψήφιο αγοραστή κατοικίας προτείνοντάς του μία ή παραπάνω κατοικίες από αυτές που υπάρχουν στη βάση δεδομένων-γνώσεων του. Συγκεκριμένα, το σύστημα θα ρωτάει από το χρήστη τις εξής πληροφορίες:
   - Αν προτιμά μικρό, μεσαίο ή μεγάλο σπίτι.
   - Σε ποια περιοχή της Θεσσαλονίκης (Ανατολικά, Δυτικά, Κέντρο).
   - Καινούργιο ή παλιό.
   - Πολυκατοικία ή μεζονέτα.
   - Πόσα μέλη έχει η οικογένεια.

- Αν έχει αυτοκίνητο.

Οι παραπάνω πληροφορίες θα αξιοποιούνται από το σύστημα γνώσης με τον ακόλουθο τρόπο:

α) Η προτίμηση σε μέγεθος σε συνάρτηση με τον αριθμό των μελών της οικογένειας καθορίζει τις απαιτήσεις σε εμβαδόν. Συγκεκριμένα, ο παρακάτω πίνακας καθορίζει τα τετραγωνικά εμβαδού που αντιστοιχούν ανά άτομο σε κάθε κατηγορία μεγέθους:

Κατηγορία	Τετραγωνικά ανά άτομο
Μικρό	μικρότερο από 25 m^2
Μεσαίο	από 25 έως 35 m^2
Μεγάλο	πάνω από 35 m^2

β) Ο αριθμός των μελών της οικογενείας καθορίζει επίσης και τον αριθμό των υπνοδωματίων που απαιτούνται από το σπίτι, επιπρόσθετα από το εμβαδόν. Συγκεκριμένα, για *n* άτομα ο μέγιστος αριθμός υπνοδωματίων είναι *n*-1, ενώ ο ελάχιστος είναι τέτοιος ώστε σε κάθε υπνοδωμάτιο να αντιστοιχούν το πολύ 2 άτομα.

γ) Η περιοχή στην οποία πρέπει να βρίσκεται στο σπίτι, καθορίζει τη συνοικία ως εξής:

Περιοχή	Συνοικία
Ανατολική	Καλαμαριά, Πυλαία, Τούμπα
Κέντρο	Άνω Πόλη, Άγιος Παύλος, Συκιές
Δυτική	Σταυρούπολη, Εύοσμος, Μενεμένη

δ) Η ύπαρξη αυτοκινήτου καθορίζει αν πρέπει να έχει parking το σπίτι ή όχι.

ε) Τα υπόλοιπα στοιχεία, δηλαδή καινούργιο-παλιό σπίτι και πολυκατοικία-μεζονέτα χρησιμοποιούνται απευθείας για επιλογή σπιτιού.

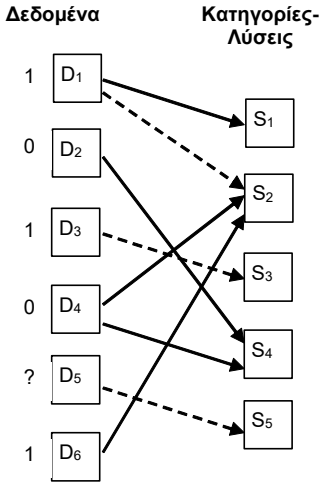
Παραδείγματα σπιτιών που θα έχει στη βάση του το σύστημα είναι τα εξής:

ID	Εμβαδόν (m^2)	Αριθμός δωματίων	Καινούργιο (Κ) Παλιό (Π)	Πολυκατοικία (Π) Μεζονέτα (Μ)	Parking (Ν/Ο)	Συνοικία	Τιμή (€)
1	95	3	Π	Π	Ο	Καλαμαριά	190,000
2	105	2	Π	Μ	Ν	Πυλαία	189,000
3	111	3	Κ	Π	Ν	Τούμπα	177,600
4	84	2	Π	Π	Ο	Άνω Πόλη	142,800
5	97	2	Κ	Μ	Ο	Άγιος Παύλος	145,500
6	93	2	Κ	Π	Ο	Συκιές	120,900
7	120	3	Κ	Μ	Ο	Σταυρούπολη	144,000
8	130	4	Κ	Π	Ν	Εύοσμος	130,000
9	92	2	Π	Π	Ο	Μενεμένη	128,800
10	115	3	Κ	Π	Ν	Πυλαία	210,000

Το σύστημα γνώσης πρέπει να δέχεται όλα τα δεδομένα από το χρήστη υπό μορφή απαντήσεων σε ερωτήσεις που υποβάλει. Τα δεδομένα των σπιτιών θα βρίσκονται προ-αποθηκευμένα ως γεγονότα ή αντικείμενα. Στο τέλος, το σύστημα θα προτείνει ένα ή περισσότερα σπίτια που ικανοποιούν τις απαιτήσεις του υποψήφιου αγοραστή, τυπώνοντάς τα στην οθόνη με ευανάγνωστο τρόπο. Αν δεν υπάρχουν

σπίτια που ταιριάζουν στον αγοραστή, τότε να τυπώνεται κάποιο μήνυμα που να το υποδηλώνει.

4. Στο μοντέλο κατηγοριοποίησης του σχήματος (συζευκτικό μοντέλου κατηγοριοποίησης), το διάνυσμα των δεδομένων εισόδου ποιες κατηγορίες εξηγεί και ποιες καλύπτει και γιατί; Τι θα άλλαζε αν το δεδομένο D_5 έπαιρνε τιμή 0 ή 1;



5. Όπως αναφέρθηκε στην αρχή του κεφαλαίου, η κατηγοριοποίηση μπορεί να χρησιμοποιηθεί για διάγνωση. Μία εταιρία χημικών προϊόντων έχει ένα πρόβλημα εντοπισμού χημικών μολύνσεων στο εργοστάσιό της. Στο εργοστάσιο διατηρούνται τα εξής χημικά:

Ελληνική ονομασία	Αγγλική ονομασία
Θειικό οξύ	Sulphuric acid
Υδροχλωρικό οξύ	Hydrochloric acid
Οξικό οξύ	Acetic acid
Ανθρακικό οξύ	Carbonic acid
Υδροξείδιο του Νατρίου	Sodium hydroxide
Χρωμογόνο 23	Chromogen 23
Υδροξείδιο του Αργιλίου	Aluminium hydroxide
Υδροξείδιο του Ρουβιδίου	Rubidium hydroxide
Βενζίνη	Petrol
Πετρέλαιο	Transformer oil

Ιδιότητες Χημικών

Τα παραπάνω χημικά ανήκουν σε μία από τις 3 κατηγορίες: οξέα (acid), βάσεις (base) και έλαια (oil). Τα οξέα έχουν pH μικρότερο από 6 και οι βάσεις έχουν pH μεγαλύτερο ή ίσο του 8. Τα έλαια έχουν ουδέτερο pH ($6 \leq pH < 8$). Τα οξέα και οι βάσεις είναι διαλυτά στο νερό, ενώ τα έλαια όχι.

Τα οξέα χωρίζονται σε ισχυρά (strong) ($pH < 3$) και ασθενή (weak) ($3 \leq pH < 6$). Ισχυρά οξέα είναι το Θειικό και το Υδροχλωρικό. Τα ισχυρά οξέα προκαλούν εγκαύματα στο εκτεθειμένο δέρμα (burn skin). Το Θειικό οξύ μπορεί να διακριθεί από το φασματοσκοπικό εντοπισμό του Θείου (S). Το Υδροχλωρικό οξύ έχει μία έντονη αποπνικτική οσμή η οποία μπορεί να προκαλέσει ασφυξία (asphyxiation). Ασθενή οξέα είναι το Οξικό οξύ , το οποίο έχει οσμή ξυδιού, και το Ανθρακικό οξύ, το οποίο μπορεί να διακριθεί από το φασματοσκοπικό εντοπισμό του Άνθρακα (C).

Ομοίως, οι βάσεις διακρίνονται σε ισχυρές (pH≥11) και ασθενείς. Ισχυρή βάση είναι το Υδροξείδιο του Νατρίου, το οποίο μπορεί να διακριθεί από το φασματοσκοπικό εντοπισμό του Νατρίου (Na). Οι ισχυρές βάσεις, όπως και τα ισχυρά οξέα, είναι διαβρωτικές.

Οι ασθενείς βάσεις είναι το Χρωμογόνο 23, το Υδροξείδιο του Αργιλίου, και το Υδροξείδιο του Ρουβιδίου. Το Χρωμογόνο 23 είναι έχει κόκκινο χρώμα, ενώ το Υδροξείδιο του Αργιλίου είναι λευκό. Το Χρωμογόνο 23 είναι ελαφρύτερο από το νερό, ενώ το Υδροξείδιο του Αργιλίου βαρύτερο. Το νερό έχει ειδικό βάρος 1.0, συνεπώς τα χημικά που είναι ελαφρύτερα-βαρύτερα έχουν ειδικό βάρος μικρότερο-μεγαλύτερο από 1.0. Το Υδροξείδιο του Ρουβιδίου μοιάζει με το Υδροξείδιο του Αργιλίου στα ακόλουθα σημεία: α) έχει υψηλό ειδικό βάρος, και β) η φασματοσκοπία εντοπίζει μέταλλο. Επιπρόσθετα, το Υδροξείδιο του Ρουβιδίου είναι ραδιενεργό.

Έλαια είναι η βενζίνη και το πετρέλαιο. Τα έλαια είναι αδιάλυτα στο νερό και η φασματοσκοπία εντοπίζει Άνθρακα. Η βενζίνη είναι ελαφρύτερη από το νερό και εμφανίζει κίνδυνο ανάφλεξης-έκρηξης (explosive). Το πετρέλαιο περιέχει ισχυρά τοξικά στοιχεία PCB (highly toxic PCBs).

Για τα χημικά που δεν έχει σαφώς αναφερθεί κάτι προηγουμένως, ισχύει ότι δεν εντοπίζεται τίποτα στη φασματοσκοπία, ότι είναι άχρωμα, άοσμα, έχουν ειδικό βάρος 1.0, και δεν είναι ραδιενεργά.

Σταθμός Ελέγχου

Ο σταθμός ελέγχου έχει τη δυνατότητα για τις εξής μετρήσεις:

Ελληνικός όρος	Αγγλικός όρος	Τιμές
pH	pH	πραγματικός αριθμός [0, 14]
διαλυτότητα	solubility	διαλυτό/αδιάλυτο (soluble/insoluble)
φασματοσκοπία	spectroscopy	Κανένα, ένα ή περισσότερα από τα ακόλουθα στοιχεία: Άνθρακας/θείο/μέταλλο/νάτριο (carbon/sulphur/metal/sodium)
χρώμα	colour	Λευκό/κόκκινο/άχρωμο (White/red/none)
οσμή	smell	Αποπνικτική/ξυδιού/άοσμο (Choking/vinegar/none)
ειδικό βάρος	specific gravity	πραγματικός αριθμός [0.9, 1.1]
ραδιενέργεια	radioactivity	υπάρχει ή όχι

Ο σταθμός ελέγχου αναφέρει μόνο τις μετρήσεις για τις οποίες είναι 100% σίγουρος, δηλαδή δεν υπάρχει αβεβαιότητα στην ορθότητα της μέτρησης. Τα δεδομένα που αναφέρονται από το σταθμό πρέπει να θεωρούνται 100% αξιόπιστα, δηλαδή όταν εμφανίζεται κάποια μόλυνση και πρέπει να αναζητηθεί το(-α) συγκεκριμένο(-α) χημικό(-ά) που προκάλεσε(-αν) τη μόλυνση, αυτό(-α) πρέπει να εξηγεί(-ούν) όλες τις αναφερθείσες μετρήσεις.

Συλλογιστική του ειδικού

Ένας ειδικός στις χημικές μολύνσεις προσπαθεί να εντοπίσει το χημικό που προκάλεσε τη μόλυνση ταυτοποιώντας (ταιριάζοντας) τις παρατηρούμενες μετρήσεις με τις γνωστές ιδιότητες των χημικών. Στην προσπάθειά του αυτή μπορεί να εντοπίσει με σιγουριά το ένα και μοναδικό χημικό που προκάλεσε τη μόλυνση, αλλά μπορεί και όχι, δηλαδή μπορεί περισσότερα του ενός χημικά να είναι υποψήφια.

Απαιτήσεις της άσκησης

Να αναπτυχθεί ένα σύστημα γνώσης σε γλώσσα ή εργαλείο της προτίμησής σας το οποίο να κωδικοποιεί όλη την παραπάνω γνώση και να επιλύει το πρόβλημα του εντοπισμού του είδους της μόλυνσης. Το σύστημα γνώσης, χρησιμοποιώντας τις πληροφορίες (μετρήσεις) που καταγράφει ο σταθμός ελέγχου, θα πρέπει να εντοπίζει το(-α) χημικό(-ά) το(-α) οποίο(-α) προκάλεσε(-αν) τη μόλυνση στο ποτάμι. Το

πρόγραμμα θα πρέπει να αναφέρει στην οθόνη τα ονόματα των χημικών, καθώς και τους πιθανούς κινδύνους που σχετίζονται με κάθε ένα από αυτά.

Το πρόγραμμα θα πρέπει να είναι σε θέση να επιλύει σωστά τουλάχιστον τις ακόλουθες περιπτώσεις, χωρίς αυτό να σημαίνει ότι δεν θα επιλύει και άλλες:

α) Η φασματοσκοπία εντοπίζει μέταλλο και υπάρχει ραδιενέργεια. *Απάντηση*: Το χημικό που προκάλεσε τη μόλυνση είναι το υδροξείδιο του ρουβιδίου.

β) Το χημικό είναι αδιάλυτο στο νερό με ειδικό βάρος 0.95. *Απάντηση*: Το χημικό που προκάλεσε τη μόλυνση είναι η βενζίνη.

γ) Το pH του χημικού είναι 2. *Απάντηση*: Τα χημικά που μπορεί να προκάλεσαν τη μόλυνση είναι το Θειικό οξύ και το Υδροχλωρικό οξύ.

6. Στο εμπόριο υπάρχουν διάφορα είδη πυροσβεστήρων που χρησιμοποιούνται σε διαφορετικές περιπτώσεις. Για παράδειγμα, σε περίπτωση φωτιάς σε ηλεκτρική συσκευή απαγορεύεται η χρήση πυροσβεστήρων που έχουν σα βάση το νερό. Τα υλικά που υπάρχουν χωρίζονται σε τέσσερις κατηγορίες: A, B, C και D. Για κάθε μια κατηγορία υπάρχουν διαφορετικά μέσα πυρόσβεσης. Στον επόμενο πίνακα δίνονται οι τέσσερις κατηγορίες και μερικά υλικά που ανήκουν σε αυτές:

A	wood, cloth, paper
B	oil, gas, grease
C	oven, TV, heater
D	sodium, potassium, magnesium

Παρακάτω φαίνονται τα υλικά που χρησιμοποιούνται για την κατάσβεση των τεσσάρων τύπων φωτιάς.

A	dry chemicals, water, waterbased liquids
B	dry chemicals, carbon dioxide, foam, bromotrifluoromenthane
C	dry chemicals, carbon dioxide, bromotrifluoromenthane
D	trimethoxyboroxine, screened graphitized coke

Να γραφεί ένα σύστημα γνώσης σε γλώσσα ή εργαλείο της προτίμησής σας το οποίο να ρωτά το χρήστη για το υλικό το οποίο καίγεται και μετά να τυπώνει στην οθόνη τα είδη πυροσβεστήρων που μπορεί αυτός να χρησιμοποιήσει, με βάση τους παραπάνω πίνακες.

7. Να γραφεί ένα σύστημα γνώσης σε γλώσσα ή εργαλείο της προτίμησής σας για την επιλογή αγοράς αυτοκινήτου. Τα αυτοκίνητα θα εξετάζονται για τα ακόλουθα χαρακτηριστικά:

- ο τύπος αυτοκινήτου (οικογενειακό, πολυτελείας, πόλης)
- η τιμή (ακριβό, φτηνό, κανονική τιμή)
- οι επιδόσεις του αυτοκινήτου (υψηλές, μέτριες, χαμηλές)
- ο εξοπλισμός (πλούσιος, φτωχός)

Επίσης ισχύουν οι εξής κανόνες:

- Ένα αυτοκίνητο θεωρείται ακριβό εάν κοστίζει περισσότερο από 20,000 €.
- Ένα αυτοκίνητο θεωρείται ότι έχει κανονική τιμή εάν κοστίζει πάνω από 10,000 € και μέχρι 20,000 €.
- Ένα αυτοκίνητο θεωρείται φθηνό εάν κοστίζει λιγότερο από 10,000 €.

Τα αυτοκίνητα τα οποία θα πρέπει να εισαχθούν μαζί με τις τιμές τους για τα χαρακτηριστικά είναι τα ακόλουθα:

Αυτοκίνητο	Τύπος	Τιμή	Επιδόσεις	Εξοπλισμός
Opel Astra	Οικογενειακό	15,000€	Χαμηλές	Πλούσιος
Peugeot 106	Πόλης	9,000€	Υψηλές	Φτωχός
Mercedes E200	Πολυτελείας	36,000€	Μέτριες	-
Rover 25	Οικογενειακό	18,000€	-	Πλούσιος
Ferrari F40	Πολυτελείας	84,000€	Υψηλές	Υψηλές

Βάσει του παραπάνω πίνακα να γραφούν κανόνες για τα αυτοκίνητα ώστε να προτείνεται το κατάλληλο αυτοκίνητο, σύμφωνα με τα χαρακτηριστικά που εισήγαγε ο χρήστης.

Παρατηρήσεις

- Όταν για κάποιο χαρακτηριστικό δε δίνεται τιμή στον πίνακα, να θεωρηθεί ότι το συγκεκριμένο αυτοκίνητο επιλέγεται ανεξάρτητα του χαρακτηριστικού αυτού.
- Για τα επιθυμητά χαρακτηριστικά του αυτοκίνητου θα ερωτάται ο χρήστης κατά τη διάρκεια εκτέλεσης του προγράμματος. Όσον αφορά την τιμή του αυτοκινήτου, θα ζητείται από το χρήστη μια τιμή μεταξύ των "ακριβό, φτηνό, κανονική τιμή".
- Το ή τα αυτοκίνητα που τελικά επιλέγονται θα εμφανίζονται είτε στην οθόνη ή στη λίστα γεγονότων.

8. Ποια είναι η κυριότερη διαφορά της μεθόδου κατηγοριοποίησης "*K2: Από τα Δεδομένα σε Πιθανές Λύσεις*" από την μέθοδο "*K1: Παραγωγή και Δοκιμή*"; Ποιες είναι οι κυριότερες ομοιότητες και διαφορές της μεθόδου κατηγοριοποίησης "*K4: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τα Δεδομένα*" από την μέθοδο "*K3: Ιεραρχική Κατηγοριοποίηση Καθοδηγούμενη από τις Λύσεις*", καθώς και από την μέθοδο *K2*;

9. Ποιες διαφοροποιήσεις εισήγαγε το σύστημα MYCIN στην "κλασική" ανάστροφη ακολουθία εκτέλεσης; Ποια γενική μέθοδο κατηγοριοποίησης από τις K1-K4 χρησιμοποιεί το σύστημα MYCIN; Δικαιολογήστε την απάντησή σας.

10. Ποια γενική μέθοδο κατηγοριοποίησης από τις K1-K4 χρησιμοποιεί το σύστημα PROSPECTOR; Δικαιολογήστε την απάντησή σας χρησιμοποιώντας τις φάσεις εκτέλεσης του συστήματος PROSPECTOR: *εισαγωγή δεδομένων*, *προώθηση πιθανοτήτων*, *επιβεβαίωση υπόθεσης*.

11. Έστω ότι θέλουμε να επιλέξουμε ένα δώρο για κάποιον φίλο μας. Σύμφωνα με τις συνήθειες μας, επιλέγουμε τα δώρα μας εξετάζοντας το προφίλ του φίλου μας και συγκεκριμένα την ηλικία του, αν είναι μορφωμένος, αν του αρέσουν τα ακριβά δώρα, και αν του αρέσει η μουσική. Συγκεκριμένα, έχουμε κατατάξει τους φίλους μας σε τρεις ηλικιακές ομάδες: μικρή (μικρότερος των 15 ετών), μεσαία (μεταξύ 15 και 34) και μεγάλη (άνω των 34), και συνήθως επιλέγουμε:
    - CD μουσικής, για κάποιον που του αρέσει η μουσική και ανήκει στην μικρή ή στην μεσαία ηλικιακή ομάδα.
    - Παιχνίδια, για κάποιον που του αρέσουν τα ακριβά δώρα και ανήκει στην μικρή ηλικιακή ομάδα.

- Ρούχα, για κάποιον που του αρέσουν τα ακριβά δώρα και ανήκει στην μεσαία ηλικιακή ομάδα.
- Λουλούδια, για κάποιον που του αρέσουν τα ακριβά δώρα και ανήκει στην μεγάλη ηλικιακή ομάδα.
- Βιβλία, για κάποιον που είναι μορφωμένος και ανήκει στην μικρή ή στην μεσαία ηλικιακή ομάδα.

Να περιγραφεί το παραπάνω πρόβλημα ως πρόβλημα κατηγοριοποίησης χρησιμοποιώντας το μοντέλο της συζευκτικής κατηγοριοποίησης και να βρεθούν τα κατάλληλα δώρα για τις ακόλουθες περιπτώσεις φίλων μας:

α) Κάποιος(-α) που είναι 19 ετών, του αρέσει η μουσική και είναι μορφωμένος.

β) Κάποιος(-α) που είναι 40 ετών και του αρέσουν τα ακριβά δώρα.

γ) Κάποιος(-α) που είναι 40 ετών, του αρέσει η μουσική και είναι μορφωμένος.

Και στις τρεις περιπτώσεις να προσδιορίσετε τις λύσεις που ταιριάζουν με τα δεδομένα, τις λύσεις που είναι συνεπείς με τα δεδομένα, καθώς και τις λύσεις που είναι ασυνεπείς με τα δεδομένα.