ΚΕΦΑΛΑΙΟ 18

Μηχανική Μάθηση

Ασκήσεις - Ερωτήσεις

1. Να αναπτυχθούν οι έννοιες *μάθηση με επίβλεψη* και *μάθηση χωρίς επίβλεψη*, αναφέροντας σε κάθε περίπτωση και μερικές χαρακτηριστικές μεθόδους.

2. Γράψτε την ακολουθία των συνόλων *S* και *G* όπως υπολογίζονται από τον αλγόριθμο απαλοιφής υποψηφίων αν οι 5 περιπτώσεις του παραδείγματος (Πίνακας 18.2) δινόταν με αντίστροφη σειρά.

3. Γράψτε την ακολουθία των συνόλων *S* και *G* όπως υπολογίζονται από τον αλγόριθμο απαλοιφής υποψηφίων σύμφωνα με τα παραδείγματα του πίνακα.

Ημέρα	Θερμ/σία Αέρα	Υγρασία	Άνεμος	Θερμ/σία Νερού	Πρόβλεψη	Θαλάσσιο Άθλημα
1	Μικρή	Κανονική	Ισχυρός	Μικρή	Σταθερός	Ναι
2	Μικρή	Υψηλή	Ισχυρός	Μικρή	Σταθερός	Ναι
3	Μεγάλη	Υψηλή	Ισχυρός	Μικρή	Αλλαγή	Όχι
4	Μικρή	Υψηλή	Ισχυρός	Μεγάλη	Αλλαγή	Ναι

4. Ποιο είναι το βασικό πλεονέκτημα του αλγόριθμου απαλοιφής υποψηφίων συγκριτικά με τον αλγόριθμο ID3;

5. Τι ονομάζονται *δένδρα ταξινόμησης (απόφασης)*; Πότε χρησιμοποιούνται σε σχέση με άλλα μοντέλα λήψης απόφασης; Πότε ένα δένδρο *ταξινόμησης* χαρακτηρίζεται ως *αμιγές* (*pure*) και γιατί τα *αμιγή δέντρα* δεν είναι επιθυμητά;

6. Προτείνετε 3 λύσεις αντιμετώπισης του προβλήματος ελλιπών τιμών στα δένδρα ταξινόμησης.

7. Ποιο είναι το κυριότερο πρόβλημα του Κέρδους Πληροφορίας στην κατασκευή δένδρων ταξινόμησης; Αναφέρετε ένα άλλο στατιστικό μέτρο που το αντιμετωπίζει. Πως μπορούμε να δώσουμε προβάδισμα σε χαρακτηριστικά με μικρότερο κόστος;

8. Έστω το παρακάτω σύνολο δεδομένων. Η εντροπία των παραδειγμάτων με τιμή Χαμηλή για το χαρακτηριστικό Ανεργία είναι 0 και η εντροπία των παραδειγμάτων με τιμή Κακός για το χαρακτηριστικό Πληθωρισμός είναι 1.

	Πληθωρισμός	**Ανεργία**	**Πορεία Μετοχής**
1	Καλός	Χαμηλή	
2	Καλός	Υψηλή	
3	Κακός	Χαμηλή	Καθοδική
4	Κακός	Υψηλή	

Επίσης, είναι γνωστό ότι ο αλγόριθμος ID3, βρίσκει την Ανεργία να είναι καταλληλότερο χαρακτηριστικό από τον Πληθωρισμό για το διαχωρισμό στη ρίζα. Οι τιμές που μπορεί να πάρει η εξαρτημένη μεταβλητή (Πορεία Μετοχής) είναι Καθοδική και Ανοδική.

Συμπληρώστε τις τιμές που λείπουν από τον πίνακα, αιτιολογώντας την απάντησή σας.

9. Γιατί κατά την ταξινόμηση είναι καλό να υπάρχουν τόσο θετικά όσο και αρνητικά παραδείγματα μιας κατηγορίας;

10. Να κατασκευαστεί αναλυτικά το δένδρο ταξινόμησης του προβλήματος δανειοδότησης (Σχήμα 18.12).

11. Υπολογίστε με βάση τα δεδομένα του πίνακα:

- Την εντροπία σε σχέση με την κατηγορία (+, -)
- Το κέρδος πληροφορίας του α_2

Περίπτωση	**Κατηγορία**	**α_1**	**α_2**
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

12. Έστω ένα πρόβλημα κατασκευής ενός δένδρου απόφασης με βάση το σύνολο *S* των εγγραφών του παρακάτω πίνακα, στον οποίο καταγράφεται το αν έγινε ένας αθλητικός αγώνας σε σχέση με τις συνθήκες υγρασίας, ανέμου και θερμοκρασίας που επικρατούσαν.

Ημέρα	Υγρασία	Άνεμος	Θερμοκρασία	Έγινε Αγώνας?
H1	υψηλή	ασθενής	υψηλή	όχι
H2	υψηλή	ισχυρός	υψηλή	όχι
H3	υψηλή	ασθενής	μέση	όχι
H4	κανονική	ασθενής	χαμηλή	ναι
H5	κανονική	ισχυρός	μέση	ναι

α) Χρησιμοποιώντας τα μεγέθη *Εντροπία* και *Κέρδος* και θεωρώντας ως εξαρτημένη μεταβλητή το πεδίο "*Έγινε Αγώνας*", να αποφασιστεί ποιο από τα πεδία υγρασία, άνεμος και θερμοκρασία είναι καταλληλότερο για τον επόμενο διαχωρισμό.
β) Να κατασκευαστεί το πλήρες δένδρο ταξινόμησης (απόφασης).

13. Έστω το παρακάτω σύνολο δεδομένων. Με βάση ποιο χαρακτηριστικό θα γίνει ο διαχωρισμός στη ρίζα, αν εφαρμόσουμε τον αλγόριθμο ID3; Για τη διακριτοποίηση του συνεχούς χαρακτηριστικού επιλέξτε το κατάλληλο κατώφλι c. Δίνονται όλοι οι λογάριθμοι που απαιτούνται.

Καιρός	Θερμοκρασία	Υγρασία	Τένις
Ηλιόλουστος	30	Χαμηλή	Όχι
Συννεφιασμένος	10	Υψηλή	Ναι
Ηλιόλουστος	20	Χαμηλή	Ναι
Βροχερός	5	Υψηλή	Όχι
Ηλιόλουστος	15	Χαμηλή	Ναι
Συννεφιασμένος	7	Υψηλή	Ναι
Συννεφιασμένος	2	Χαμηλή	Όχι
Βροχερός	29	Υψηλή	Όχι

14. Έστω το παρακάτω σύνολο δεδομένων.

- Με βάση ποιο χαρακτηριστικό (Φύλο ή Ύψος) θα γίνει ο διαχωρισμός στη ρίζα, αν εφαρμόσουμε τον αλγόριθμο ID3;
- Για τη διακριτοποίηση του συνεχούς χαρακτηριστικού επιλέξτε το κατάλληλο(α) κατώφλι(α) με τη μέθοδο του διαχωρισμού ίσης συχνότητας.

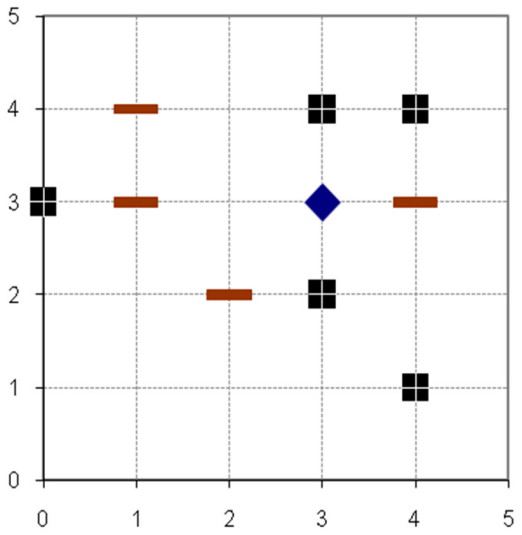
ID	Φύλλο	Ύψος	Χαρακτηρισμός
1	Θ	1.6	Κοντός/ή
2	Θ	1.6	Κοντός/ή
3	Θ	1.7	Κοντός/ή
4	Α	1.7	Κοντός/ή
5	Θ	1.75	Μέτριος/α
6	Θ	1.8	Μέτριος/α
7	Θ	1.8	Μέτριος/α
8	Α	1.85	Μέτριος/α
9	Θ	1.88	Μέτριος/α
10	Θ	1.9	Μέτριος/α
11	Θ	1.9	Μέτριος/α
12	Α	1.95	Μέτριος/α
13	Α	2.0	Ψηλός/ή
14	Α	2.1	Ψηλός/ή
15	Α	2.2	Ψηλός/ή

15. Να χρησιμοποιήσετε το dataset Breast cancer και να εφαρμόσετε έναν classification tree αλγόριθμο δοκιμάζοντας διάφορες τιμές των παραμέτρων: split function (criterion) και maxdepth. Επίσης να εφαρμόσετε τον random forest με τιμές παραμέτρων split function (criterion) και number of trees (n_estimators).

    Τα αποτελέσματα θα εμφανίζονται σε έναν πίνακα όπου να αναγράφονται οι μετρικές Precision, Recall και F1 για κάθε περίπτωση. Στο τέλος να δώστε μια σύντομη παράγραφο σχολιασμού των αποτελεσμάτων.

16. Να υπολογιστεί η τιμή του *x*' στο Σχήμα 18.14 χρησιμοποιώντας τη μέθοδο των *k-κοντινότερων γειτόνων* και λαμβάνοντας διαδοχικά 2, 3, 4, 6, 7 και 8 κοντινότερους γείτονες.

17. Να ταξινομηθεί το στιγμιότυπο (3,3) του σχήματος: α) Με τον αλγόριθμο 5- πλησιέστερων γειτόνων, β) Με τον αλγόριθμο πλησιέστερων γειτόνων σταθμισμένης απόστασης. Να χρησιμοποιήσετε την απόσταση Manhattan και στα 2 ερωτήματα.



18. Για τα δεδομένα του παρακάτω πίνακα, προβλέψτε την τιμή της κατηγορίας (Καλός ή Κακός πελάτης) για την περίπτωση 6 με τη χρήση του αλγορίθμου κοντινότερων γειτόνων (KNN) σταθμισμένης απόστασης.

    Χρησιμοποιήστε την απόσταση Manhattan.(Τα ποσά στα εισοδήματα/οφειλές είναι σε χιλιάδες euro).

Πελάτης	Εισοδήματα	Οφειλές	Κατηγορία
1	15	65	Κακός
2	20	40	Κακός
3	40	20	Καλός
4	60	30	Καλός
5	40	40	Καλός
6	30	40	?

19. Για τα δεδομένα του παρακάτω πίνακα, προβλέψτε την άγνωστη περίπτωση <100, West, 5> με τη χρήση του αλγορίθμου 5 κοντινότερων γειτόνων σταθμισμένης απόστασης:

α/α	Μέγεθος (τμ)	Περιοχή	Όροφος	Τιμή (Κ €)
1	100	East	2	200
2	120	East	3	250
3	80	East	4	150
4	90	Centre	2	250
5	100	Centre	3	280
6	120	Centre	4	360
7	110	West	2	100
8	130	West	3	120
9	90	West	4	90
10	150	West	2	120

20. Ποιο είναι το σημαντικότερο μειονέκτημα του ταξινομητή Bayes. Πως αντιμετωπίζεται;

21. Έστω το παρακάτω σύνολο δεδομένων.

    - Πώς θα κατηγοριοποιηθεί η νέα περίπτωση (Μέτριο, 30) σύμφωνα με τον αφελή ταξινομητή Bayes;
    - Διακριτοποιήστε το χαρακτηριστικό Ηλικία με τη μέθοδο του διαχωρισμού ίσης συχνότητας σε δύο κλάσεις (Μικρή/Μεγάλη Ηλικία).

Παράδειγμα	Εισόδημα	Ηλικία	Έκδοση Πιστωτικής
1	Υψηλό	34	Ναι
2	Μέτριο	42	Ναι
3	Μέτριο	46	Ναι
4	Υψηλό	40	Ναι
5	Χαμηλό	27	Όχι
6	Υψηλό	23	Όχι
7	Χαμηλό	53	Όχι
8	Μέτριο	26	Όχι

22. Ποιο θα θεωρούσατε ως το κυριότερο πρόβλημα της μάθησης με Νευρωνικά Δίκτυα;

23. Να αναφέρετε δυο πλεονεκτήματα και δυο μειονεκτήματα των Μηχανών Διανυσμάτων Υποστήριξης (ΜΔΥ)

24. Έστω το παρακάτω σύνολο δεδομένων. Αν η εξαρτημένη μεταβλητή είναι η Ανεργία, διακριτοποιήστε τη μεταβλητή Πληθωρισμός σε δύο τιμές (χαμηλός, υψηλός) με τη μέθοδο της εντροπίας. (Δεν χρειάζεται να γίνουν πράξεις)

Δημόσια έσοδα	Πληθωρισμός	Ανεργία
Μικρά	2	Σταθερή
Μεσαία	3	Πτωτική
Μεγάλα	4	Πτωτική

25. Έστω δύο δυαδικοί ταξινομητές Α και Β με αποτελέσματα αξιολόγησης που φαίνονται στους παρακάτω πίνακες ενδεχομένων ή σύγχυσης (confusion matrices).

A	**Προβλεπόμενη Κλάση**		
Πραγματική Κλάση		**Ναι**	**Όχι**
	Ναι	80	20
	Όχι	30	70
B	**Προβλεπόμενη Κλάση**		
Πραγματική Κλάση		**Ναι**	**Όχι**
	Ναι	90	10
	Όχι	40	60

- Απεικονίστε τους ταξινομητές στο χώρο ROC.
- Ποιόν ταξινομητή θα επιλέγατε με βάση την ανάλυση ROC;
- Ποιον θα επιλέγατε με βάση την ακρίβεια (ACC)
- Ποιον θα επιλέγατε με βάση το F-measure

26. Πώς αξιολογείται η ποιότητα των κανόνων συσχέτισης; Ποιες είναι οι διαφορές και ποιες οι ομοιότητες των κανόνων συσχέτισης με τους κανόνες ταξινόμησης;

27. Έστω το συχνό σύνολο αντικειμένων $\{A,B,C\}$ με υποστήριξη (support) 0.15. Δίνονται: *Support*(A)=0.3, *Support*(B)=0.3, *Support*(C)=0.4, *Support*($\{A,B\}$)= 0.25, *Support*($\{A,C\}$)=0.2, *Support*($\{B,C\}$)=0.3. Βρείτε τους κανόνες με την υψηλότερη τιμή εμπιστοσύνης που παράγονται από αυτό το σύνολο.

28. Με τον όρο γονιδιακή έκφραση ή έκφραση γονιδίων (gene expression) στην γενετική χαρακτηρίζεται η διαδικασία εκείνη που προκαλεί τη μεταφορά κωδικοποιημένων πληροφοριών (του γονιδίου) στο λειτουργικό προϊόν του γονιδίου(πρωτεΐνη ή RNA). Στον παρακάτω πίνακα συναλλαγών φαίνεται η έκφραση τεσσάρων γονιδίων (Γi). Κάθε συναλλαγή αντιπροσωπεύει μια διαφορετική μέτρηση των επιπέδων έκφρασης του κάθε γονιδίου. Όλες οι μετρήσεις έχουν διακριτοποιηθεί έτσι ώστε να υπάρχουν μόνο δύο επίπεδα έκφρασης (1 ή 0) που αντιστοιχεί στο αν εκφράζεται ένα γονίδιο ή όχι.

Μέτρηση	**Γ_1**	**Γ_2**	**Γ_3**	**Γ_4**
T_1	1	1	1	0
T_2	0	0	0	1
T_3	1	0	1	0
T_4	1	1	1	1
T_5	0	1	0	0

Οι κανόνες που θα μπορούσαν να προκύψουν είναι της μορφής:

«Αν εκφραστεί το γονίδιο1 τότε εκφράζεται το γονίδιο3»

Να υπολογίσετε την υποστήριξη, την εμπιστοσύνη και το Lift για τους παρακάτω κανόνες:

$\{\Gamma_1\} \Rightarrow \{\Gamma_2\}$

$\{\Gamma_1, \Gamma_2\} \Rightarrow \{\Gamma_3\}$

29. Ένα κατάστημα γυναικείων ενδυμάτων έχει 20 συναλλαγές στο ταμείο κατά τη διάρκεια μιας ημέρας, όπως φαίνεται στον παρακάτω πίνακα:

Συναλλαγή	Στοιχεία
T1	Blouse
T2	Shoes, Skirt, TShirt
T3	Jeans, TShirt
T4	Jeans, Shoes, TShirt
T5	Jeans, Shorts
T6	Shoes, TShirt
T7	Jeans, Skirt
T8	Jeans, Shoes, Shorts, TShirt
T9	Jeans
T10	Jeans, Shoes, TShirt
T11	TShirt
T12	Blouse, Jeans, Shoes, Skirt, TShirt
T13	Jeans, Shoes, Shorts, TShirt
T14	Shoes, Skirt, TShirt
T15	Jeans, Tshirt
T16	Skirt, TShirt
T17	Blouse, Jeans, Skirt
T18	Jeans, Shoes, Shorts, TShirt
T19	Jeans
T20	Jeans, Shoes, Shorts, TShirt

Να υπολογίσετε την υποστήριξη (support), την εμπιστοσύνη (confidence) και την άνωση (Lift) για τους παρακάτω κανόνες

{Shoes} ⇒{TShirt}

{Shoes} ⇒ {TShirt, Jeans}

30. Σε ένα πρόβλημα ταξινόμησης ο ταξινομητής που δημιουργήθηκε είχε την ακόλουθη επίδοση στα δεδομένα ελέγχου:

α/α	1	2	3	4	5	6	7	8	9	10	11	12	13	14
γνωστή κλάση	Όχι	Όχι	Ναι	Ναι	Ναι	Όχι	Ναι	Όχι	Ναι	Ναι	Ναι	Ναι	Ναι	Όχι
πρόβλεψη	Όχι	Όχι	Όχι	Ναι	Ναι	Όχι	Όχι	Όχι	Ναι	Ναι	Όχι	Ναι	Ναι	Ναι

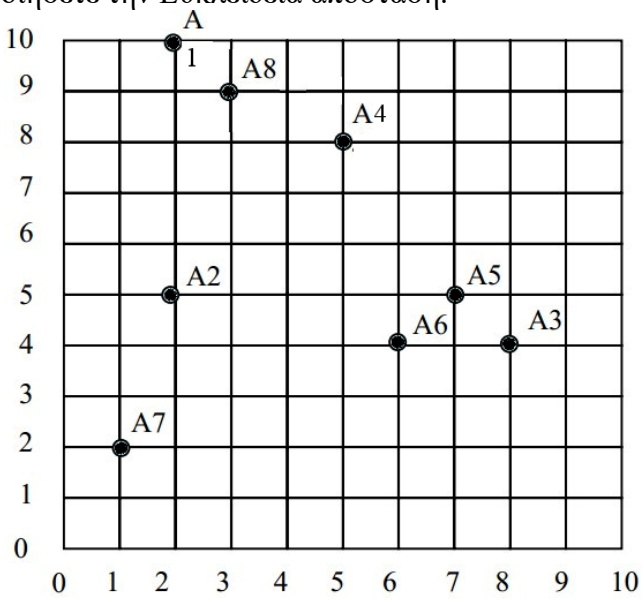
Να δημιουργήσετε τον πίνακα σύγχυσης με βάση το διάγραμμα. Στη συνέχεια να υπολογίσετε Recall, Precision και Accuracy. Στον πίνακα να φαίνεται ξεκάθαρα σε ποια πλευρά είναι η πρόβλεψη καθώς και το τι μετράει κάθε κουτάκι (π.χ. για True Positive να γράψετε στο σχετικό κουτάκι: TP=12)

31. Ποιες είναι οι παράμετροι αξιολόγησης της ποιότητας μιας ομαδοποίησης και τι υπολογίζει η κάθε μία; Ποιος είναι ο συντελεστής αξιολόγησης μιας ομάδας ή μιας ομαδοποίησης;

32. Έστω ένα σύνολο σημείων που παρουσιάζεται στο παρακάτω σχήμα.

α) Εφαρμόστε μία επανάληψη του αλγόριθμου των 3-μέσων για να ομαδοποιήσετε όλα τα σημεία. Ως αρχικά κέντρα για τις τρεις ομάδες να θεωρήσετε τα σημεία A1, A4 και A7. Για τον υπολογισμό των αποστάσεων να χρησιμοποιήσετε την απόσταση Μανχάταν.

β) Ποια είναι η απόσταση μεταξύ της ομάδας που περιέχει το A1 και της ομάδας που περιέχει το A7 μετά την εκτέλεση μιας επανάληψης του αλγορίθμου; Υπολογίστε την απόσταση με βάση τα κέντρα των ομάδων. Σε αυτή την περίπτωση να χρησιμοποιήσετε την Ευκλείδεια απόσταση.



33. Κάποιος σας ζητάει να κατασκευάσετε ένα σύστημα το οποίο έπειτα από την εισαγωγή των στοιχείων ενός πελάτη θα πληροφορεί το χειριστή του συστήματος εάν ο πελάτης είναι καλός ή όχι. Οι απαιτήσεις του συστήματος είναι οι εξής: μικρός χρόνος απόκρισης και εύκολη ερμηνεία της διαδικασίας απόφασης σχετικά με τον πελάτη.

α) Ποιους από τους παρακάτω αλγορίθμους μηχανικής μάθησης θα επιλέγατε για να υλοποιήσετε το σύστημα:
   1. K-μέσων
   2. K- Πλησιέστερων Γειτόνων
   3. C4.5
   4. SVM
   5. NN

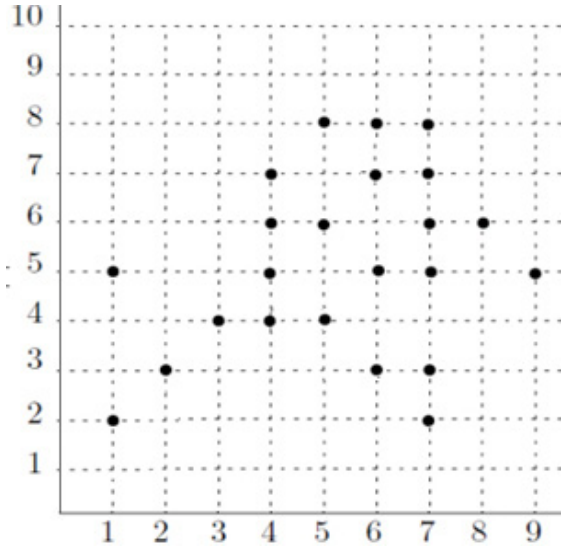
   Αιτιολογήστε γιατί επιλέξατε ή απορρίψατε τον κάθε αλγόριθμο.

β) Αν η μοναδική απαίτηση του συστήματος ήταν η ικανότητα του να προστίθενται εύκολα στη γνώση του οι καινούριες περιπτώσεις πελατών, ποιον από τους παραπάνω αλγορίθμους θα επιλέγατε ως καταλληλότερο; Γιατί;

γ) Τι θα κάνατε για να αυξήσετε την ακρίβεια ταξινόμησης;

34. Έχετε τα δεδομένα από συμβόλαια κινητής τηλεφωνίας κάποιας εταιρίας. Πώς μπορεί η συσταδοποίηση να δώσει αξία σε αυτά τα δεδομένα ώστε να ωφεληθεί η εταιρία;

35. Στο σχήμα, αν η ακτίνα eps=1 και το MinPts=3, να κυκλώσετε τα σημεία πυρήνα και να βάλετε X στα συνοριακά σημεία. Όσα απομένουν θα θεωρηθούν σημεία θορύβου. Στη συνέχεια να ορίσετε τις ομάδες (clusters) περικυκλώνοντας με μια γραμμή τα σημεία κάθε ομάδας.



36. Σε ένα πρόβλημα ομαδοποίησης μονοδιάστατων δεδομένων, τα σημεία προς ομαδοποίηση έχουν τις ακόλουθες συντεταγμένες (βλ. X άξονα): A=0.05, B=0.15, C=0.6, D=1, E=0.38, F=0.42. Να εφαρμόσετε ιεραρχική ομαδοποίηση μέχρι να καταλήξετε σε μία ομάδα. Σε κάθε νέα υπο-ομάδα που προκύπτει να υπολογίζετε τις συντεταγμένες του κέντρου της και να το απεικονίζετε στη σωστή θέση στο διάγραμμα. Να εξηγήσετε αναλυτικά μία περίπτωση συγχώνευσης, δηλαδή πώς προέκυψαν οι συντεταγμένες του νέου κέντρου.