Στο έγγραφο αυτό συγκεντρώνουμε τις ποιό συχνές παρατηρήσεις που κάναμε στην πρόοδο της εξόρυξης γνώσης από δεδομένα.

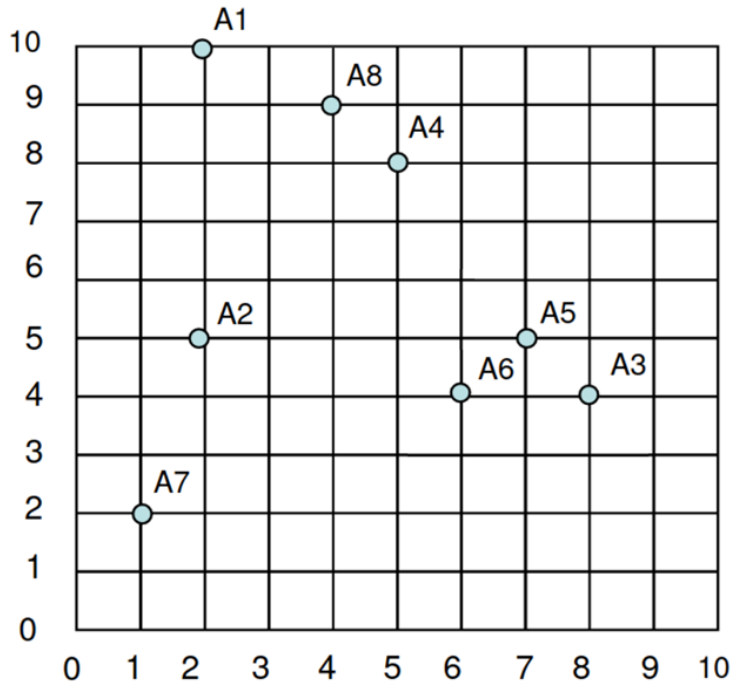
Σημειώστε βέβαια ότι σε όλες τις ερωτήσεις οι μονάδες δεν ήταν "όλες ή μηδέν", αλλά προσμετρήθηκαν κλιμακωτά, ανάλογα την απάντηση.

Ποιο ή ποιά από τα ακόλουθα είναι παραδείγματα εξαγωγής χαρακτηριστικών;

Α. Κατασκευή διανύσματος "bag of words" από ένα email
Β. Εφαρμογή PCA σε μεγάλα δεδομένα υψηλής διαστάσεων
Γ. Αφαίρεση stopwords από μια πρόταση
Δ. Καμία από τις υπόλοιπες απαντήσεις

Α, Β και Γ.

Δίνονται τα 8 σημεία (A1 ως A8) στον δισδιάστατο χώρο, τα οποία τα συσταδοποιούμε με τη χρήση του αλγορίθμου DBSCAN με MinPts=2 και Eps=√10 . Να γράψετε ποιες συστάδες σχηματίζονται, αν υπάρχουν οριακά σημεία (border points), ποια είναι και σε ποια συστάδα είναι οριακά καθώς και αν υπάρχουν σημεία θορύβου (noise points) και ποια είναι αυτά.



Εφόσον MinPts =2 και Eps = 3,16, αρχικά χρειάζεται να χαρακτηρίσουμε τα σημεία μας σε σχέση με τον 2ο πλησιέστερο γείτονα τους και ακτινας = 3.16 . Τα A5,A6,A3 είναι εντός μιας συστάδας γιατι η αποσταση πχ του A5 απο το A6 ειναι sqrt((7-6)^2+(5-4)^2) = dqrt(2)<eps. Αντίστοιχα προκύπτει και για το A5-A3 και εφόσον αυτά τα τρία ειναι διπλα το ένα στο άλλο για καθένα από αυτά αποτελούν 2ο πλησιέστερο γείτονα τα αλλα δυο. Η δευτερη συστάδα αποτελείται απο τα A8,A4 και A1. Η τελευταία συστάδα είναι η A2,A7.

Με βάση τα παραπάνω,τα σημεία Α5,Α2 και Α8 αποτελούν σημεία πυρήνα και το Α7 αποτελεί οριακό σημείο γιατί απέχει απόσταση από το Α2 ακριβώς απόσταση ριζα(10).

Έχουμε έξι δείγματα δύο χαρακτηριστικών το καθένα: [4 1], [6 6], [9 5], [1 2], [7 3], [5 4]. Οι ετικέτες τους είναι [1 0 1 0 1 0]. Για την ταξινόμηση φτιάχνουμε ένα δέντρο αποφάσεων βάθους δύο. Οι διχοτομήσεις προκύπτουν με κανόνες που θέτουν ένα χαρακτηριστικό μεγαλύτερο ίσο από ένα κατώφλι. α) ποια είναι η εντροπία της ρίζας του δέντρου; β) ποιος είναι ο κανόνας της δεύτερης διχοτόμησης;

α) $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$ ή απλά αν οι ετικέτες είναι σε δυαδικό πρόβλημα οι ετικέτες είναι μισές-μισές η εντροπία είναι μέγιστη δηλαδή 1.

β) Κατατάσσουμε τα δείγματα και τις ετικέτες με βάση το πρώτο χαρακτηριστικό x1
1 2 **0**
4 1 **1**
5 4 **0**
6 6 **0**
7 3 **1**
9 5 **1**
Κάνουμε το ίδιο για το δεύτερο χαρακτηριστικό x2
4 **1 1**
1 **2 0**
7 **3 1**
5 **4 0**
9 **5 1**
6 **6 0**
Ο καλύτερος πρώτος κανόνας είναι x1>=7.
Για τη δεύτερη διχοτόμηση: το υποδέντρο με ετικέτες [1, 1] δεν χρειάζεται να χωριστεί.
Κάνουμε τις δύο κατατάξεις για το υποδέντρο με ετικέτες [0, 1, 0, 0]
1 2 **0**
4 1 **1**
5 5 **0**
6 5 **0**
και
4 **1 1**
1 **2 0**
5 **4 0**
6 **6 0**
Ο καλύτερος δεύτερος κανόνας είναι ο x2>=2

Έχουμε έξι δείγματα δύο χαρακτηριστικών: [4 1], [6 6], [9 5], [1 2], [7 3], [5 4]. Οι ετικέτες τους είναι [1 0 1 0 1 0]. Για την ταξινόμηση φτιάχνουμε ένα δέντρο βάθους δύο. Οι διχοτομήσεις προκύπτουν με κανόνες που θέτουν ένα χαρακτηριστικό μεγαλύτερο ίσο από ένα κατώφλι. α) στη ρίζα του δέντρου, ποιοι κανόνες δίνουν μηδενικό κέρδος πληροφορίας; β) ποιος είναι ο κανόνας της πρώτης διχοτόμησης;

α) Κατατάσσουμε τα δείγματα και τις ετικέτες με βάση το πρώτο χαρακτηριστικό x1

1 2 **0**
4 1 **1**
5 4 **0**
6 6 **0**
7 3 **1**
9 5 **1**
Κάνουμε το ίδιο για το δεύτερο χαρακτηριστικό x2
4 **1 1**
1 **2 0**
7 **3 1**
5 **4 0**
9 **5 1**
6 **6 0**
Οι κανόνες x1>=5, x2>=3 ή x2>=5 δεν πετυχαίνουν να μειώσουν την εντροπία.

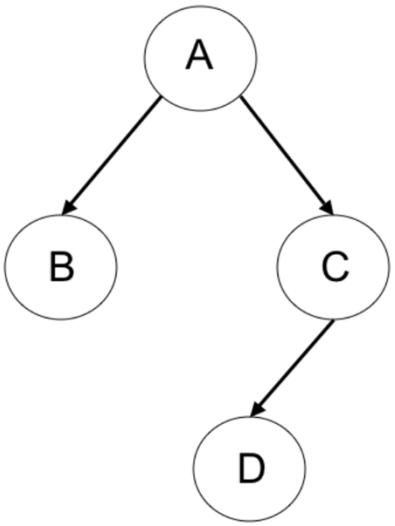
β) Ο καλύτερος πρώτος κανόνας είναι ο x1>=7.

Εκτελούμε τον αλγόριθμο Expectation-Maximization για ένα μοντέλο μίξης δύο γκαουσιανών κατανομών με ίδια βάρη για ένα σύνολο 3 δειγμάτων δεδομένων και σε κάποιο βήμα της επανάληψης έχουμε p(x1=-2|θ1)=A, p(x1=-2|θ2)=B, p(x2=1|θ1)=C, p(x2=1|θ2)=D, p(x3=3|θ1)=E, p(x3=3|θ2)=F. Έστω A=0.3, B=0.7, C=0.45, D=0.55, E=0.2, και F=0.8. Να υπολογίσετε την πιθανότητα το κάθε δείγμα δεδομένων να προέρχεται από κάθε μια από τις δύο κατανομές καθώς και τις μέσες τιμές των δύο κατανομών μετά την ολοκλήρωση αυτού του βήματος (στο διάστημα [0,1], με ακρίβεια 2 δεκαδικών ψηφίων)

p(j=1|x1=-2)=0.3, p(j=2|x1=-2)=0.7, p(j=1|x2=1)=0.45, p(j=2|x2=1)=0.55, p(j=1|x3=3)=0.2, p(j=2|x3=3)=0.8
μ1=0.47, μ2=0.76

Ενδεικτικά. Σε όλα τα versions του, το ερώτημα για τον ΕΜ μοιάζει να δυσκόλεψε.

Δίνεται το Μπεϋζιανό δίκτυο πεποίθησης του σχήματος και οι εξείς πιθανότητες: P(A)=0.3, P(B|A)=0.7, P(B|~A)=0.5, P(C|A)=0.2, P(C|~A)=0.6, P(D|C)=0.7, P(D|~C)=0.7. Υπολογίστε την πιθανότητα P(C|B)



P(C|B)=P(C,B)/P(B) Ομως p(B,C) = P(B,C|A)+P(B,C|~A) = P(B|A)P(C|A) P(A)+P(B|~A)P(C|~A) P(~A)=0.7x0.2x0.3+0.6x0.5x0.7 = 0.252 P(B) = P(B|A)P(A)+P(B|~A)P(~A) = 0.7x0.3+0.5x0.7 = 0.56 Αρα P(C|B) = 0.45

Ενδεικτικά. Σε όλα τα versions του, το ερώτημα για τα δίκτυα πεποίθησης μοιάζει να δυσκόλεψε.