1ο Σύνολο Ασκήσεων
Ημερομηνία Παράδοσης: 8 Απριλίου 2008, στο μάθημα
Ενότητα: Συσταδοποίηση

Οι ασκήσεις με χαρακτηρισμό **Α** είναι ατομικές, ενώ οι ασκήσεις με χαρακτηρισμό **Δ** μπορεί να γίνουν σε ομάδες έως 2 ατόμων.

Οι ασκήσεις με **(*)** είναι προαιρετικές με την παρακάτω έννοια: μπορείτε να τις παραδώσετε αντί τελικής εξέτασης. Θα υπάρχουν αντίστοιχες και στα άλλα σύνολα. Για αυτούς που θα επιλέξουν να ολοκληρώσουν όλες τις ασκήσεις (δηλαδή και τις προαιρετικές), ο τελικός βαθμός τους θα προκύψει από τον βαθμό τους στα σύνολα ασκήσεων. Για τους υπόλοιπους, οι ασκήσεις θα συμμετέχουν με ποσοστό 50% στον τελικό τους βαθμό.

Ποσοστό επί του τελικού βαθμού: 30 % για όσους ασχοληθούν με τις ασκήσεις (*)
15 % για τους υπόλοιπους

Για τους αλγόριθμους συσταδοποίησης, μπορείτε να χρησιμοποιήστε τα εργαλεία WEKA, MATLAB είτε δικό σας κώδικα (είτε αν θέλετε κάποιο άλλο εργαλείο). Πληροφορίες για τα εργαλεία WEKA και MATLAB υπάρχουν στην ιστοσελίδα του μαθήματος.

Άσκηση 1 [Δ, ()]
Υλοποιείστε μια απλή εκδοχή του k-mean για δυσ-διάστατα δεδομένα στο [0, 10] που θα χρησιμοποιεί την ευκλείδεια απόσταση.
(α) Δημιουργείστε 500 τυχαία σημεία και τρέξτε τον αλγόριθμο με k = 10.
(β) Δημιουργείστε 500 σημεία που να ανήκουν σε 10 κύκλους με την ίδια ακτίνα και ξένους μεταξύ τους. Αναθέστε (περίπου) ίδιο αριθμό σημείων σε κάθε κύκλο και τρέξτε τον αλγόριθμο με k = 5, k = 10 και k = 20. Δώστε μια επιλογή αρχικών σημείων που να οδηγεί σε καλά αποτελέσματα και μια που δεν οδηγεί.
(γ) Υπολογίστε τη συνεκτικότητα και το διαχωρισμό για τα ερωτήματα (α) και (β) με k = 10.
(δ) Αναπαραστήσετε το αποτέλεσμα του αλγόριθμου σας για τα ερωτήματα (α) και (β), τυπώνοντας τα σημεία χρησιμοποιώντας διαφορετικό σύμβολο ή χρώμα για κάθε συστάδα.

Άσκηση 2 [Δ, (*)]
Εφαρμόστε ιεραρχική συσταδοποίηση στα δεδομένα της Άσκησης 1(β).
Δώστε το δεντρό-γραμμα που προκύπτει από τον συσσωρευτικό αλγόριθμο ιεραρχικής συσταδοποίησης χρησιμοποιώντας: (α) MIN, (β) MAX και (γ) μέσο όρο.

Άσκηση 3 [Δ, (*)]
Τρέξτε τον αλγόριθμο DBSCAN στα δεδομένα της Άσκησης 1(β). Πειραματιστείτε με την επιλογή του Eps και MinPts. Δώστε 3 διαφορετικές συσταδοποιήσης για διαφορετικές επιλογές αυτών των τιμών.

Άσκηση 4 [Α, ()]
Για το σύνολο της Άσκησης 1(β), εξηγείστε πως θα επιλέγατε τιμές για το Eps και MinPts χρησιμοποιώντας τη μέθοδο που δώσαμε στο μάθημα.

Άσκηση 5 [Α, ( )]
Θεωρείστε τα παρακάτω 6 στοιχεία (εγγραφές). Κάθε στοιχείο έχει 5 δυαδικά γνωρίσματα. : A(10110), B(11011), C(10110), D(01010), E(10101) και F(01110) και τον παρακάτω πίνακα ομοιότητας (contingency) μεταξύ δύο στοιχείων R και S.

	1	0
1	M_{11}	M_{10}
0	M_{01}	M_{00}

Θεωρείστε τις παρακάτω συναρτήσεις ομοιότητας:
$SMC(R, S) = (M_{11} + M_{00}) / (M_{11} + M_{10} + M_{01} + M_{00})$ $J(R, S) = (M_{11}) / (M_{11} + M_{10} + M_{01})$ και $RC(R, S) = (M_{11}) / (M_{11} + M_{10} + M_{01} + M_{00})$.
Χρησιμοποείστε ιεραρχικό αλγόριθμο συσταδοποίησης για να δώσετε τα δεντρο-γράμματα για τις παρακάτω περιπτώσεις: (α) απόσταση MIN και SMC , (β) απόσταση MAX και J, και (γ) απόσταση μέσο όρο και RC