1ο Σύνολο Ασκήσεων
Καταληκτική Ημερομηνία Παράδοσης: 30 Απριλίου 2009, στις 17:00, στο Εργαστήριο Κατανεμημένων (B15)
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-means για δυσ-διάστατα δεδομένα στο [0, 10] που θα χρησιμοποιεί (i) την Ευκλείδεια απόσταση (L2) και μια εκδοχή που θα χρησιμοποιεί (ii) την L1 (Manhattan ή city-block) απόσταση και θα επιλέγει ως κεντρικό σημείο κάθε συστάδα το μεσαίο σημείο (όχι το αριθμητικό μέσο, όπως στην Ευκλείδεια απόσταση).
(α) Δημιουργείστε 300 τυχαία σημεία και τρέξτε τον αλγόριθμο με k = 10.
(β) Δημιουργείστε 300 σημεία που να ανήκουν σε 6 κύκλους με την ίδια ακτίνα και ξένους μεταξύ τους. Αναθέστε (περίπου) ίδιο αριθμό σημείων σε κάθε κύκλο και τρέξτε τον αλγόριθμο με k = 3, k = 6 και k = 12. Δώστε μια επιλογή αρχικών σημείων που να οδηγεί σε καλά αποτελέσματα και μια που δεν οδηγεί.
(γ) Υπολογίστε τη συνεκτικότητα και το διαχωρισμό για το ερώτημα (β) με k = 3 και k =6 για την Ευκλείδεια απόσταση. Σε ποια από τις δυο περιπτώσεις, η συσταδοποίηση είναι καλύτερη.
(δ) Αναπαραστήσετε το αποτέλεσμα του αλγόριθμου σας για τα ερωτήματα (α) και (β), τυπώνοντας τα σημεία χρησιμοποιώντας διαφορετικό σύμβολο ή χρώμα για κάθε συστάδα.
(ε) (προαιρετικά) Εξηγείστε τη διαφορετική επιλογή κεντρικού σημείου για την L1.

Άσκηση 2 [Δ, (*)]
Εφαρμόστε ιεραρχική συσταδοποίηση στα δεδομένα της Άσκησης 1(β), θεωρώντας Ευκλείδεια απόσταση.
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), θεωρώντας Ευκλείδεια απόσταση. Πειραματιστείτε με την επιλογή του Eps και MinPts. Δώστε 2 διαφορετικές συσταδοποιήσεις 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, ( )]
(α) Θεωρείστε ένα σύνολο από σημεία τέτοια ώστε τα σημεία να ανήκουν σε τρεις φυσικές συστάδες, για το οποία ο k-means πιθανότατα θα έβρισκε τις σωστές συστάδες, αλλά ο k-means με διχοτόμηση θα αποτύγχανε.
(β) Εξηγείστε το ρόλο του κατωφλιού T στον αλγόριθμο Birch.