1ο Σύνολο Ασκήσεων
Ημερομηνία Παράδοσης: 13 Απριλίου 2011, στο μάθημα
Ενότητα: Συσταδοποίηση, Κατηγοριοποίηση

Ποσοστό του τελικού βαθμού: 40 % του ως απαλλακτικές
20 % αν δώσετε τελικό διαγώνισμα

ΟΔΗΓΙΕΣ

Οι ασκήσεις 1, 2 και 3 αφορούν πειραματισμό με τις μεθόδους εξόρυξης δεδομένων που καλύψαμε στο μάθημα και μπορούν να παραδοθούν και σε ομάδες των 2 ατόμων.
Για τις ασκήσεις αυτές μπορείτε να χρησιμοποιήστε το εργαλείο WEKA, τις αντίστοιχες υλοποιήσεις σε MATLAB είτε δικό σας κώδικα (είτε αν θέλετε κάποιο άλλο εργαλείο). Κάποιες πληροφορίες για τα εργαλεία WEKA και MATLAB υπάρχουν και στην ιστοσελίδα του μαθήματος.

Οι ασκήσεις 4, 5 και 6 είναι θεωρητικές και ατομικές.

Οι ασκήσεις είναι απαλλακτικές, με την έννοια ότι μπορεί να αναπληρώσουν το τελικό διαγώνισμα (δείτε και τη σελίδα του μαθήματος).

Άσκηση 1 [σε ομάδες έως 2 ατόμων] [30 μονάδες]

1. Γράψτε ένα πρόγραμμα που θα παράγει N 2-διάστατα σημεία που να ανήκουν σε m κύκλους με την ίδια ακτίνα. Ο αριθμός των σημείων N, ο αριθμός m και η ακτίνα των κύκλων θα είναι είσοδος στο πρόγραμμά σας.

2. Χρησιμοποιείστε το πρόγραμμά σας για να δημιουργείστε 500 σημεία που να ανήκουν σε 5 ξένους μεταξύ τους κύκλους. Αναθέστε (περίπου) τον ίδιο αριθμό σημείων σε κάθε κύκλο.

Στα παρακάτω ερωτήματα θα πειραματιστείτε με τον k-means.

(α) Τρέξτε τον k-means με k = 5, 10, 20 και Ευκλείδεια (L2) απόσταση.
(β) Τρέξτε τον k-means με k = 5, 10, 20 και Manhattan (L1) απόσταση.
(γ) Τρέξτε τον k-means με k = 20 και Ευκλείδεια (L1) απόσταση πολλές φορές.

Για καθένα από τα παραπάνω, αναπαραστήστε το αποτέλεσμα τυπώνοντας στην οθόνη τα σημεία χρησιμοποιώντας διαφορετικό σύμβολο ή χρώμα για κάθε συστάδα. Στόχος είναι να δείτε τα αποτελέσματα του k-means.

3. Χρησιμοποιείστε το πρόγραμμά σας για να δημιουργείστε 500 σημεία που να ανήκουν σε 5 κύκλους με την ίδια ακτίνα και εφαπτομένους μεταξύ τους. Αναθέστε (περίπου) τον ίδιο αριθμό σημείων σε κάθε κύκλο.

Στα παρακάτω ερωτήματα θα πειραματιστείτε με τον συσωρευτικό αλγόριθμο ιεραρχικής συσταδοποίησης.

(α) Τρέξτε τον αλγόριθμο χρησιμοποιώντας MIN (single link) απόσταση,
(β) Τρέξτε τον αλγόριθμο χρησιμοποιώντας MAX (complete link) απόσταση,
(γ) Τρέξτε τον αλγόριθμο χρησιμοποιώντας average (μέσο όρο) απόσταση,

Για καθένα από τα παραπάνω, αναπαραστήστε το αποτέλεσμα τυπώνοντας στην οθόνη τα σημεία χρησιμοποιώντας διαφορετικό σύμβολο ή χρώμα για κάθε συστάδα. Δείτε επίσης το δενδρόγραμμα. Στόχος είναι να δείτε τα αποτελέσματα του ιεραρχικού αλγορίθμου.

4. Χρησιμοποιείστε το πρόγραμμά σας για να δημιουργείστε 500 σημεία που να ανήκουν σε 3 εφαπτόμενους κύκλους.

Στα παρακάτω ερωτήματα θα πειραματιστείτε με τον αλγόριθμο DBScan.

(α) Αναθέστε (περίπου) τον ίδιο αριθμό σημείων σε κάθε κύκλο. Τρέξτε τον αλγόριθμο DBScan. Πειραματιστείτε με την επιλογή του Eps και MinPts.
(β) Αναθέστε σε έναν από τους 3 κύκλους διπλάσιο αριθμό σημείων από ότι στους άλλους δύο. Τρέξτε τον αλγόριθμο DBScan. Πειραματιστείτε με την επιλογή του Eps και MinPts ώστε να έχετε ως αποτέλεσμα τον ίδιο αλλά και διαφορετικό αριθμό από συστάδες από ότι στο ερώτημα (α).

Για καθένα από τα παραπάνω, αναπαραστήστε το αποτέλεσμα τυπώνοντας στην οθόνη τα σημεία χρησιμοποιώντας διαφορετικό σύμβολο ή χρώμα για κάθε συστάδα. Στόχος είναι να δείτε τα αποτελέσματα του DBScan.

5. Για κάθε μία από τις παραπάνω περιπτώσεις συσταδοποίησης (δηλαδή, για τα ερωτήματα 2, 3 και 4) προσθέστε θόρυβο στα δεδομένα σας (δηλαδή, έναν αριθμό σημείων εκτός των κύκλων) και ξανατρέξτε τους αλγόριθμους.

Τι θα παραδώστε:
Θα υπάρξει προφορική εξέταση που θα μας δείξετε τα αποτελέσματα των συσταδοποιήσεων.
Για την άσκηση αυτή, θα παραδώστε μόνο απαντήσεις στα παρακάτω:
1. Για τον k-means (ερώτημα 2):
(α) Εξηγείστε τη διαφορά στα αποτελέσματα για k=5, 10, 20
(β) Εξηγείστε τη διαφορά στα αποτελέσματα για Ευκλείδεια και Manhattan
(γ) Εξηγείστε γιατί υπάρχει διαφορά ανάμεσα στα πολλαπλά τρεξίματα
(δ) Υπολογίστε τη συνοχή και το διαχωρισμό για τις συσταδοποιήσεις με k = 5 και k = 10 και Ευκλείδεια απόσταση. Ποια είναι καλύτερη;
2. Για την ιεραρχική συσταδοποίηση (ερώτημα 3), εξηγείστε τα αποτελέσματα για τις διαφορετικές αποστάσεις.
3. Για το DBScan (ερώτημα 4), εξηγείστε την επιλογή των Eps και MinPts με βάση τη μέθοδο με τους γείτονες που εξηγήσαμε στο μάθημα.
4. Για το θόρυβο (ερώτημα 5), εξηγείστε ποιοι αλγόριθμοι επηρεάζονται από το θόρυβο και ποιοι όχι και γιατί.

Άσκηση 2 [σε ομάδες έως 2 ατόμων] [15 μονάδες]

Σκοπός της άσκησης είναι η εξοικείωσή σας με ένα εργαλείο για κατηγοριοποίηση με χρήση δέντρων απόφασης. Αν χρησιμοποιείστε WEKA, χρησιμοποιείστε το J48 (υλοποιεί τον C4.5).

1. Τρέξτε τον αλγόριθμο για τα δεδομένα weather θεωρώντας ως γνώρισμα «κλάση» (κατηγορία) το play. Χρησιμοποιείστε 10 cross-validation. Δώστε το δέντρο που προκύπτει.

2. Επαναλάβετε το ερώτημα 1 χρησιμοποιώντας (α) 66% των δεδομένων ως δεδομένα εκπαίδευσης και 33% ως δεδομένου ελέγχου και (β) 33% των δεδομένων ως δεδομένα εκπαίδευσης και 66% ως δεδομένα ελέγχου.

3. Τρέξτε τον αλγόριθμο για τα δεδομένα zoo θεωρώντας ως γνώρισμα «κλάση» (κατηγορία) το type. Χρησιμοποιείστε 10 cross-validation. Δώστε το δέντρο που προκύπτει.

Τι θα παραδώστε:
Θα υπάρξει προφορική εξέταση που θα μας δείξετε τα αποτελέσματα της κατηγοριοποίησης.
Για την άσκηση αυτή, θα παραδώστε μόνο απαντήσεις στα παρακάτω:
1. Εξηγείστε για όλα τα ερωτήματα τις τιμές των παραμέτρων εισόδου
2. Συγκρίνετε τα αποτελέσματα του ερωτήματος 2(α) και 2(β) και σχολιάστε τη διαφορά τους.

Άσκηση 3 [σε ομάδες έως 2 ατόμων] [10 μονάδες]

Σκοπός της άσκησης είναι η εξοικείωσή σας με ένα εργαλείο για κατηγοριοποίηση με χρήση k-κοντινότερων γειτόνων. Μπορείτε να χρησιμοποιείστε Matlab.

1. Επιλέξτε ένα σύνολο δεδομένων (dataset) για κατηγοριοποίηση από το UCI repository.
2. Χωρίστε το σύνολο δεδομένων σε σύνολα εκπαίδευσης και ελέγχου και τρέξτε τον αλγόριθμο για τέσσερις διαφορετικές του K

Τι θα παραδώστε: Τι θα παραδώστε:
Θα υπάρξει προφορική εξέταση που θα μας δείξετε τα αποτελέσματα της κατηγοριοποίησης.
Για την άσκηση αυτή, θα παραδώστε μόνο απαντήσεις στα παρακάτω:
1. Εξηγείστε ποιο σύνολο δεδομένων διαλέξατε και πως το χωρίσατε.

2. Δώστε μια αξιολόγηση των αποτελεσμάτων της κατηγοριοποίησης για τις διαφορετικές τιμές του k με βάση το ρυθμό σφάλματος (ή κάποιο άλλο κατάλληλο μέτρο).

Άσκηση 4 [ατομική] [15 μονάδες]

Έστω τα παρακάτω δυαδικά διανύσματα:
(1, 0, 1, 1, 0), (1, 1, 0, 1, 1), (1, 0, 1, 1, 0), (0, 1, 0, 1, 0), (1, 0, 1, 0, 1) και (0, 1, 1, 1, 0).
Χρησιμοποιείστε ιεραρχικό αλγόριθμο συσταδοποίησης και δώστε τα δενδρογράμματα για τις παρακάτω περιπτώσεις: (α) απόσταση MIN και συντελεστή απλού ταιριάσματος (SMC) και (β) απόσταση MAX και συντελεστή Jaccard (τους ορισμούς των συντελεστών μπορείτε να τους βρείτε στο κεφάλαιο 2.4.5)

Άσκηση 5 [ατομική] [20 μονάδες]

Θεωρείστε τα παρακάτω δεδομένα (Πίνακας 1) για ένα δυαδικό πρόβλημα κατηγοριοποίησης.

Πίνακας 1. Σύνολο Δεδομένων για την Άσκηση 4

A	**B**	**Κατηγορία**
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

1. Κατασκευάστε το δέντρο απόφασης χρησιμοποιώντας εντροπία. Υπολογίστε το πεσιμιστικό σφάλμα εκπαίδευσης (με όρο ποινής ίσο με 0.5). Κατασκευάστε τον πίνακα σύγχυσης και υπολογίστε την πιστότητα (accuracy), ανάκληση (recall) και ακρίβεια (precision) για την κατηγορία +.

2. Κατασκευάστε το δέντρο απόφασης χρησιμοποιώντας το ευρετήριο Gini. Πως εξηγείτε το γεγονός ότι επιλέγεται διαφορετικό γνώρισμα από ότι στο ερώτημα 1, αν και το ευρετήριο Gini και η εντροπία είναι μονότονα αύξουσες στο [0, 0.5] και φθίνουσες στο [0.5, 1];

3. Χωρίστε τα δεδομένα σε δύο σύνολα – οι 7 πρώτες εγγραφές στο σύνολο εκπαίδευσης και οι υπόλοιπες στο σύνολο ελέγχου. Κατασκευάστε το δέντρο απόφασης χρησιμοποιώντας λάθος κατηγοριοποίησης. Υπολογίστε το σφάλμα γενίκευσης και το F1.

Άσκηση 6 [ατομική] [10 μονάδες]

Στον αλγόριθμο BIRCH:

1. Εξηγείστε το ρόλο του κατωφλίου T, και

2. Δείξτε πως μπορεί να υπολογιστεί η απόσταση ενός σημείου από μια συστάδα με βάση την περίληψη της συστάδας (clustering feature)