2ο Σύνολο Ασκήσεων

Ημερομηνία Παράδοσης: 16 Μαΐου 2008 (μέχρι τις 13μμ στη βοηθό του μαθήματος)
Ενότητα: Κανόνες Συσχέτισης, Ταξινόμηση

Ποσοστό επί του τελικού βαθμού: **45 %** για όσους ασχοληθούν με τις ασκήσεις (*)
23 % για τους υπόλοιπους

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.

Για τους αλγόριθμους, μπορείτε να χρησιμοποιήστε τα εργαλεία WEKA, MATLAB, δικό σας κώδικα, ή κάποιο άλλο εργαλείο. Πληροφορίες για τα εργαλεία WEKA και MATLAB υπάρχουν στην ιστοσελίδα του μαθήματος.

Άσκηση 1 [Α, ()]

(α) Έστω ότι χωρίζουμε ένα σύνολο δοσοληψιών D σε n ξένα μεταξύ τους υποσύνολα. Δείξτε ότι ένα στοιχειοσύνολο που είναι συχνό στο D πρέπει να είναι συχνό σε τουλάχιστον ένα από τα n υποσύνολα.
(β) Θεωρείστε ότι η βάση δεδομένων των δοσοληψιών D είναι κατανεμημένη σε m κόμβους.. Δηλαδή, κάθε κόμβος αποθηκεύει τοπικά ένα (διαφορετικό) υποσύνολο των δοσοληψιών. Χρησιμοποιείστε την ιδιότητα που αποδείξατε στο (α) για να σχεδιάστε μια παραλλαγή του apriori αλγορίθμου υπολογισμού συχνών στοιχειοσυνόλων που να αποφεύγει να μεταφέρει όλες τις δοσοληψίες σε έναν μόνο κόμβο και έτσι να είναι αποδοτικός από πλευράς κόστους επικοινωνίας.
(γ) Έστω ότι έχουμε ένα σύνολο δοσοληψιών D για το οποίο έχουμε υπολογίσει και αποθηκεύσει τα συχνά στοιχειοσύνολα με ελάχιστη υποστήριξη (misup) s. Έστω ότι προσθέτουμε δ καινούργιες δοσοληψίες στο D. Συζητείστε πως μπορούμε να βρούμε με αποδοτικό τρόπο τα συχνά στοιχειοσύνολα με το ίδιο s στο νέο σύνολο που περιλαμβάνει και τις καινούργιες δοσοληψίες (χωρίς να τρέξουμε από την αρχή τον αλγόριθμο στο νέο σύνολο).

Άσκηση 2 [Α, (*)]

Θεωρείστε τον παρακάτω σύνολο δοσοληψιών και ελάχιστη υποστήριξη 4 (minsup = 40%).

TID	**Στοιχεία**
T10	{6, 1, 3}
T20	{1, 2, 4, 5, 3}
T30	{3, 2, 5}
T40	{6, 7}
T50	{1, 3, 2, 4, 5}
T60	{1, 3, 6}
T70	{1, 2, 5, 7}
T80	{2, 8, 5, 1}
T90	{4, 6}
T100	{1, 2, 5}

(α) Εφαρμόστε τον αλγόριθμο FP-Growth για να βρείτε τα συχνά στοιχειοσύνολα. Δώστε το αρχικό FP-δέντρο (θεωρείστε τα στοιχεία ταξινομημένα με βάση τη συχνότητα εμφάνισής τους), τα προθεματικά δέντρα καθώς και τα συχνά στοιχειοσύνολα και την υποστήριξη τους που προκύπτουν σε κάθε βήμα.
(β) Από τα συχνά στοιχειοσύνολα που υπολογίσατε στο (α) ποια είναι maximal και ποια είναι κλειστά;
(γ) Από τα στοιχειοσύνολα που έχετε βρει στο (α) δώστε τους κανόνες με εμπιστοσύνη 100%. Για να υπολογίστε την εμπιστοσύνη κάθε κανόνα χρησιμοποιείστε την υποστήριξη που υπολογίσατε στο (α)
(δ) Δείξτε τη σειρά με την οποία ο a-priori παράγει τα συχνά στοιχειοσύνολα που υπολογίσατε στο (α).

Άσκηση 3 [Α, (*)]

Θεωρείστε τα παρακάτω συχνά 3-στοιχειοσύνολα:

{a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {b, c, d}, {b, c, e}, {c, d, e}

και ότι υπάρχουν μόνο 5 στοιχεία συνολικά..

(α) Δώστε όλα τα υποψήφια 4-στοιχειοσύνολα που δίνει η στρατηγική F_{k-1} x F_{k-1}. Στη συνέχεια, δώστε ποια από αυτά μπορούμε να απαλείψουμε/ψαλιδίσουμε (pruned).

(β) Δώστε όλα τα υποψήφια 4-στοιχειοσύνολα που δίνει η στρατηγική F_{k-1} x F_1. Στη συνέχεια, δώστε ποια από αυτά μπορούμε να απαλείψουμε/ψαλιδίσουμε (pruned).

Άσκηση 4 [Δ, ()]

Σκοπός της άσκησης είναι η εξοικείωσή σας με ένα εργαλείο για εξόρυξη κανόνων συσχέτισης. Μπορείτε να χρησιμοποιείστε το εργαλείο WEKA που υλοποιεί τον αλγόριθμο apriori.

Το εργαλείο WEKA υποστηρίζει κανόνες μόνον σε ordinal γνωρίσματα, για αυτό τον λόγο, αριθμητικά δεδομένα θα χρειαστούν προ-επεξεργασία (χρησιμοποιείστε ένα κατάλληλο φίλτρο από αυτά που είναι διαθέσιμα στο `Filter`).

(α) Εξηγείστε τι σημαίνει κάθε παράμετρος εισόδου του αλγορίθμου. (πχ στην περίπτωση της WEKA, οι παράμετροι `MetricType`, `minMetric` κοκ)

(β) Τρέξτε τον αλγόριθμο κανόνων συσχέτισης στα σύνολα δεδομένων mushroom, weather-nominal και weather (είναι τα ίδια δεδομένα όπως τα weather-nominal αλλά με αριθμητικές τιμές, για να τα χρησιμοποιείστε πρέπει να τα φιλτράρετε, χρησιμοποιείστε ένα κατάλληλο φίλτρο που να μην παράγει όμως ακριβώς τα ίδια δεδομένα με το weather-nominal). Τα σύνολα δεδομένων θα τα βρείτε στη σελίδα του μαθήματος.

Για κάθε μία περίπτωση:

(i) εξηγείστε την επιλογή των τιμών που δώσατε στις παραμέτρους εισόδου

(ii) διαλέξτε 3 από τους κανόνες εξόδου. Εκτιμήστε το ενδιαφέρον τους με βάση κάποιες από τις μετρικές που συζητήσαμε στο μάθημα και τις σχετικές μετρικές του εργαλείου (πχ στην περίπτωση της WEKA, `lift`, `convinction`, κοκ)

Άσκηση 5 [Α, (*)]

(α) Για τα δεδομένα του παρακάτω πίνακα υπολογίστε το ευρετήριο Gini

(i) για το γνώρισμα Φύλο

(ii) για το γνώρισμα Είδος Αυοκινήτου με πολλαπλό διαχωρισμό

(iii) για το γνώρισμα Είδος Μέγεθος Ρούχων με πολλαπλό διαχωρισμό

(β) Ποιο θα επιλέγατε ως γνώρισμα διαχωρισμού; Κατασκευάστε τον Πίνακα Confusion για το δέντρο απόφασης που προκύπτει από αυτήν την επιλογή. Υπολογίστε την πιστότητα (accuracy), ανάκληση (recall) και ακρίβεια (precision).

CustomerID	Φύλο	Είδος Αυτοκινήτου	Μέγεθος Ρούχων	Κλάση
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Άσκηση 6 [Δ, ()]

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

(α) Τρέξτε τον αλγόριθμο για τα δεδομένα mushrooms θεωρώντας 2 κλάσεις: poisonous και edible. Χρησιμοποιείστε 10 cross-validation. Δώστε το δέντρο που προκύπτει.

(β) Επαναλάβετε το (α) τώρα χρησιμοποιώντας (i) 66% των δεδομένων ως δεδομένα εκπαίδευσης και 33% ως δεδομένου ελέγχου και (ii) 33% των δεδομένων ως δεδομένα εκπαίδευσης και 66% ως δεδομένου ελέγχου.

(γ) Συγκρίνετε τα δέντρα που προκύπτουν για το (α) και β(i) και β(ii) χρησιμοποιώντας κάποια από το μέτρα που μελετήσαμε στο μάθημα.

Στα (α) και (β) εξηγείστε τις τιμές των παραμέτρων.