

Πρώτη Σειρά Ασκήσεων

Οι ερωτήσεις 1-3 πρέπει να παραδοθούν στην αρχή του μαθήματος της 4^{ης} Απριλίου. Οι ερωτήσεις 4,5 θα πρέπει να γίνουν turn-in μέχρι τις 11:59 μμ την Παρασκευή 6 Απριλίου. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Οι λεπτομέρειες για το turn-in θα αναρτηθούν στη σελίδα του μαθήματος.

Άσκηση 1 (15 μονάδες)

1. Ας υποθέσουμε ότι x και y είναι δύο διανύσματα με μήκος (L_2 norm) ίσο με 1. Ποια είναι η σχέση μεταξύ της Ευκλείδειας απόστασης $d(x, y)$ και της cosine ομοιότητας $\cos(x, y)$ των δύο διανυσμάτων?
2. Ας υποθέσουμε ότι $X = (x_1, x_2, \dots, x_n)$ και $Y = (y_1, y_2, \dots, y_n)$ είναι δύο διανύσματα μεγέθους n . Το sample Pearson correlation coefficient ορίζεται ως:

$$\rho = \frac{\sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_X)^2} \sqrt{\sum_{i=1}^n (y_i - \mu_Y)^2}}$$

όπου μ_X, μ_Y είναι η μέση τιμή του διανύσματος X και Y αντίστοιχα.

Περιγράψτε πως μπορούμε να χρησιμοποιήσουμε το cosine similarity για να υπολογίσουμε το sample Pearson correlation coefficient.

3. Έστω $U = \{u_1, \dots, u_n\}$ ένα σύνολο από n αντικείμενα, και R μία διάταξη (ιεράρχηση -- ranking) των αντικειμένων. Ορίστε μια «λογική» συνάρτηση απόστασης $d(R_1, R_2)$ μεταξύ δύο διατάξεων των στοιχείων στο U . Αποδείξτε ότι η συνάρτησή σας ορίζει μία μετρική, ή το αντίθετο.

Άσκηση 2 (20 μονάδες)

Ένας χαρακτηριστικός κανόνας είναι ένας κανόνας της μορφής $\{q\} \rightarrow \{p_1, p_2, \dots, p_n\}$, όπου το αριστερό μέρος αποτελείται από ένα μόνο στοιχείο. Για ένα στοιχειοσύνολο με k στοιχεία, μπορούμε να δημιουργήσουμε k τέτοιους χαρακτηριστικούς κανόνες. Ορίζουμε το μέτρο ξ για ένα στοιχειοσύνολο, ως το ελάχιστο confidence c αυτών των κανόνων. Δηλαδή:

$$\xi(p_1, p_2, \dots, p_k) = \min\{c(\{p_1\} \rightarrow \{p_2, \dots, p_k\}), c(\{p_2\} \rightarrow \{p_1, \dots, p_k\}), \dots, c(\{p_k\} \rightarrow \{p_1, \dots, p_{k-1}\})\}$$

Είναι το μέτρο ξ μονότονο, αντί-μονότονο, ή μη μονότονο, (ούτε μονότονο, ούτε αντί-μονότονο)?

Υπενθυμίζουμε ότι για οποιαδήποτε δύο σύνολα X και Y , έτσι ώστε $X \subseteq Y$, ένα μέτρο f είναι μονότονο αν $f(X) \leq f(Y)$, και αντί-μονότονο αν $f(X) \geq f(Y)$.

Άσκηση 3 (20 μονάδες)

Υποθέστε ότι τα δεδομένα μας αποτελούνται από μία ακολουθία γεγονότων (π.χ., λέξεις σε ένα κείμενο, ή η ακολουθία κινήσεων σε ένα παιχνίδι). Θέλουμε να βρούμε όλες τις συχνές υπο-

ακολουθίες γεγονότων στα δεδομένα μας, που συμβαίνουν μέσα σε ένα χρονικό παράθυρο μεγέθους W . Για παράδειγμα, $\langle A, B, C \rangle$ είναι μια υπο-ακολουθία μήκους 3, η οποία εμφανίζεται σε ένα παράθυρο μεγέθους $W = 5$ το οποίο περιέχει τα γεγονότα $\langle A, A, B, D, C \rangle$. Για την ακρίβεια υπάρχουν δύο εμφανίσεις της υπο-ακολουθίας σε αυτό το παράθυρο, μία που ξεκινάει στην πρώτη θέση και μία που ξεκινάει στη δεύτερη θέση. Το support μιας υπο-ακολουθίας είναι ο αριθμός των παραθύρων μεγέθους W στα οποία εμφανίζεται μέσα στην συνολική ακολουθία γεγονότων.

Περιγράψτε πως θα τροποποιήσετε τον *a priori* αλγόριθμο, ώστε να υπολογίζει όλες τις συχνές υπο-ακολουθίες με support τουλάχιστον *minsup*, για δεδομένο μέγεθος παραθύρου W . Περιγράψτε ξεκάθαρα τη δημιουργία υποψήφιας υπο-ακολουθίας, το κλάδεμα των υποψηφίων, και τον υπολογισμό της συχνότητας. Μπορείτε να τροποποιήσετε τον αλγόριθμο όσο θέλετε, αρκεί να είναι σωστός και αποτελεσματικός.

Εφαρμόστε τον αλγόριθμο σας στην ακολουθία ABCACDBAAC με $W=4$ και *minsup* = 2. Αναφέρετε την έξοδο όλων των ενδιάμεσων βημάτων: υποψήφιας υπο-ακολουθίας, υπο-ακολουθίας που κλαδεύτηκαν, συχνές υπο-ακολουθίες, όλα με τη σειρά που δημιουργούνται. Αναφέρετε αν μια υπο-ακολουθία είναι το αποτέλεσμα της σύμπτυξης δύο μικρότερων ακολουθιών.

Άσκηση 4 (15 μονάδες)

Ο σκοπός αυτής της άσκησης είναι να εξοικειωθείτε με το λογισμικό εξόρυξης δεδομένων WEKA. Μπορείτε να κατεβάσετε το WEKA από τη σελίδα <http://www.cs.waikato.ac.nz/ml/weka/>. Εκεί θα βρείτε και πληροφορίες για το πώς να εγκαταστήσετε και να τρέξετε το WEKA. Πληροφορίες υπάρχουν και στη σελίδα του μαθήματος. Θα χρησιμοποιήσετε το “Congressional Voting Records” dataset (από το UCI repository), το οποίο περιέχει τις ψήφους 465 μελών του κοινοβουλίου των ΗΠΑ σε 16 διαφορετικά θέματα, και μια μεταβλητή κλάσης (δημοκρατικός, ή ρεπουμπλικάνος). Μπορείτε να βρείτε το ARFF αρχείο (η είσοδος για το WEKA) στη σελίδα του μαθήματος.

Χρησιμοποιήστε κατώφλι 0.1 για το support και βρείτε όλα τα συχνά στοιχειοσύνολα στα δεδομένα. Επίσης βρείτε τους 10 κανόνες συσχέτισης με το μεγαλύτερο confidence, και τους 10 κανόνες με το μεγαλύτερο lift (interest). Πρέπει να κάνετε turn in ένα αρχείο με τα αποτελέσματα και τις παραμέτρους που χρησιμοποιήσατε στο WEKA.

Άσκηση 5 (30 μονάδες)

Σε αυτή την άσκηση θα υλοποιήσετε το min-hashing και το Locality Sensitive Hashing. Θα τεστάρετε την υλοποίησή σας στο MovieLens 100k dataset το οποίο αποτελείται από 943 χρήστες που έχουν βαθμολογήσει 1682 ταινίες. Μπορείτε να κατεβάσετε τα δεδομένα από τη σελίδα <http://www.grouplens.org/node/73>. Διαβάστε το Readme αρχείο για λεπτομέρειες και επεξεργαστείτε τα δεδομένα όπως χρειάζεστε. Για την άσκηση μας ενδιαφέρει μόνο το σύνολο από ταινίες που βαθμολόγησε ο χρήστης και όχι οι βαθμοί. Θέλουμε να υπολογίσουμε τη Jaccard ομοιότητα μεταξύ των χρηστών.

Υπολογίστε την ακριβή Jaccard ομοιότητα μεταξύ όλων των ζευγαριών χρηστών, και τυπώστε τα ζεύγη με ομοιότητα τουλάχιστον 0.5. Μετά υπολογίστε τα min-hash signatures για όλους τους χρήστες και την

προσεγγιστική Jaccard ομοιότητα όπως περιγράψαμε στην τάξη. Χρησιμοποιήστε 50, 100, και 200 hash functions. Για κάθε τιμή, τυπώστε τα ζεύγη με προσεγγιστική ομοιότητα τουλάχιστον 0.5, και τον αριθμό των false positives και false negatives. Για τα false positives και false negatives αναφέρετε τη μέση τιμή από 5 διαφορετικά τρεξίματα.

Μετά, σπάστε τον πίνακα με τα signatures σε b bands με r hash functions ανά band, όπως περιγράψαμε στην τάξη, και υλοποιήστε το Locality Sensitive Hashing. Ο στόχος είναι να βρούμε υποψήφια ζεύγη με ομοιότητα τουλάχιστον 0.6. Πειραματιστείτε με $r=5$, $b=10$ για τον πίνακα με τα 50 hash functions, $r=10$, $b=10$ για τον πίνακα με τα 100 hash functions, και $r=20$, $b=10$ για τον πίνακα με τα 200 hash functions. Αναφέρετε τον αριθμό των false positives και false negatives παίρνοντας την μέση τιμή σε 5 τρεξίματα. Πως αλλάζουν αυτά τα νούμερα αν έχουμε όριο ομοιότητας 0.8?

Κάνετε turn in τον κώδικα σας, το αρχείο εισόδου που χρησιμοποιήσατε, και τα αρχεία εξόδου για τα διαφορετικά τρεξίματα. Επίσης κάνετε turn-in ένα αρχείο με τις μέσες τιμές και τις παρατηρήσεις σας.

Τεχνικές Λεπτομέρειες

1. Για το pre-processing των δεδομένων, οι παρακάτω unix εντολές μπορεί να σας φανούν χρήσιμες:
 - a. cut: επιτρέπει να πάρουμε συγκεκριμένες κολώνες από ένα αρχείο με διαχωριζόμενες τιμές
 - b. sort: ταξινομεί τις γραμμές ενός αρχείου σε αλφαβητική σειρά. -n for αριθμητική σειρά
 - c. uniq: αφαιρεί συνεχόμενες γραμμές που είναι ίδιες.
2. Χρησιμοποιείτε την παρακάτω hash function για τα signatures:
Διαλέξτε ένα αρκετά μεγάλο πρώτο αριθμό R (π.χ., $R = 131071$);
Διαλέξτε a, b, c , τυχαίους αριθμούς στο διάστημα $[0, R]$
$$h(x) = (a * (x \gg 4) + b * x + c) \% R;$$
3. Για την υλοποίηση του LSH δεν χρειάζεται να υλοποιήσετε δικό σας hash table ή list. Χρησιμοποιείτε υπάρχουσες υλοποιήσεις που έρχονται με τη γλώσσα που χρησιμοποιείτε (π.χ., στην C++ υπάρχει η STL υλοποίηση, κοιτάξτε την σελίδα του μαθήματος του αντικειμενοστρεφή προγραμματισμού για λεπτομέρειες)