Πανεπιστήμιο Ιωαννίνων,
Τμήμα Πληροφορικής

Εξόρυξη Δεδομένων – ΠΛΕ 059
Χειμερινό Εξάμηνο, 2012

Τρίτη Σειρά Ασκήσεων

Η προθεσμία για αυτή τη σειρά είναι Παρασκευή 7 Δεκεμβρίου. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Οι λεπτομέρειες για το turn-in θα αναρτηθούν στη σελίδα του μαθήματος.

Άσκηση 1

Στο βιβλίο του μαθήματος (Εισαγωγή στην Εξόρυξη Δεδομένων, των Tan, Steinbach, και Kumar) εκτός από τους αλγόριθμους clustering που περιγράψαμε στην τάξη, περιγράφονται και οι αλγόριθμοι CLARANS, BIRCH, ROCK, CHAMELEON, DENCLUE και CURE (περιγράφεται επίσης στο δωρεάν online βιβλίο Mining Massive Datasets των Rajaraman και Ullman). Διαλέξτε ένα από τους παραπάνω αλγορίθμους και περιγράψτε την κεντρική ιδέα του με δικά σας λόγια σε 2-3 παραγράφους. Μπορείτε να διαβάσετε και την σχετική δημοσίευση αν χρειάζεστε περισσότερες λεπτομέρειες.

Άσκηση 2

Αυτή είναι η Άσκηση 8.27 από το βιβλίο Εισαγωγή στην Εξόρυξη Δεδομένων, των Tan, Steinbach, και Kumar. Αποδείξτε ότι το άθροισμα των τετραγώνων των λαθών (Sum of Square Errors – SSE) ενός cluster C_i είναι ανάλογο προς το άθροισμα των τετραγώνων των αποστάσεων μεταξύ των σημείων στο cluster C_i. Πιο συγκεκριμένα, αποδείξτε ότι:

$$\sum_{x \in C_i} \|\boldsymbol{x} - \boldsymbol{c_i}\|^2 = \frac{1}{2m_i} \sum_{\boldsymbol{x} \in C_i} \sum_{\boldsymbol{y} \in C_i} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

όπου: m_i είναι ο αριθμός των σημείων στο cluster C_i ; $\boldsymbol{c_i}$ είναι το κέντρο του cluster C_i; τα σημεία στο cluster είναι d-διάστατα πραγματικά διανύσματα; και $\|\cdot\|$ είναι η Ευκλείδεια απόσταση. (**Υπόδειξη**: Ξεκινήστε με την περίπτωση όπου $d = 1$).

Άσκηση 3

Έχουμε μια συλλογή D από N κείμενα, όπου το κάθε κείμενο είναι μια «σακούλα από λέξεις» διαλεγμένες από κάποιο λεξικό W με m λέξεις. Το κείμενο d_i μπορεί να αναπαρασταθεί ως ένα m-διάστατο διάνυσμα, όπου d_{ij} είναι ο αριθμός των φορών όπου η λέξη w_j εμφανίζεται στο κείμενο d_i. Θεωρείστε ένα clustering $C = \{c_1 \dots, c_K\}$ των κειμένων στη συλλογή D, όπου $1 \leq K \leq N$. Όταν $K = N$ κάθε κείμενο είναι σε ένα cluster μόνο του (singleton clusters), ενώ όταν $K = 1$, όλα τα κείμενα είναι μαζί σε ένα cluster. Ένα cluster c_i είναι η συνένωση όλων των λέξεων από όλα τα κείμενα στο the cluster. Συνεπώς μπορούμε επίσης να το αναπαραστήσουμε ως ένα m-διάστατο διάνυσμα: το άθροισμα των διανυσμάτων των κειμένων μέσα στο cluster. Για το cluster c_i χρησιμοποιούμε n_{ij} για

τον αριθμό των εμφανίσεων της λέξης w_j στο cluster. Χρησιμοποιούμε n_i για τον αριθμό των λέξεων στο cluster c_i, και n για τον αριθμό όλων των λέξεων σε όλα τα κείμενα. Αν κανονικοποιήσουμε το διάνυσμα για το cluster c_i παίρνουμε μια κατανομή $P(W|c_i)$, όπου $p_{ij} = p(w_j|c_i) = \frac{n_{ij}}{n_i}$ είναι η δεσμευμένη πιθανότητα της λέξης w_j στο cluster c_i. Είναι η πιθανότητα αν διαλέξουμε τυχαία μια λέξη από το clusterc_i η λέξη αυτή να είναι w_j. Χρησιμοποιούμε P_i για να συμβολίσουμε την κατανομή $P(W|c_i)$. Επίσης, ορίζουμε $p(c_i) = \frac{n_i}{n}$ την πιθανότητα του cluster c_i. Είναι η πιθανότητα ότι αν διαλέξουμε τυχαία μια λέξη από όλες τις λέξεις που εμφανίζονται σε όλα τα κείμενα, η λέξη αυτή θα είναι από το cluster c_i.

Μπορούμε τώρα να ορίσουμε τη δεσμευμένη εντροπία των λέξεων W δεδομένου του clustering C. Έχουμε:

$$H(W|C) = \sum_{c_i \in C} p(c_i) H(W|c_i) = -\sum_{c_i \in C} p(c_i) \sum_{w_j \in W} p(w_j|c_i) \log p(w_j|c_i)$$

Η εντροπία των λέξεων στο cluster c_i είναι $H(W|c_i) = H(P_i)$. Ονομάζουμε την εντροπία $H(P_i)$ την *εντροπία του cluster* c_i, και την εντροπία $H(W|C)$, *την εντροπία του clustering* C.

Ας θεωρήσουμε τώρα δύο clusters (π.χ., cluster c_1 και cluster c_2), τα οποία συνενώνουμε για να δημιουργήσουμε ένα νέο cluster c_*. Έχουμε ότι $p(c_*) = p(c_1) + p(c_2)$

1. Δείξτε ότι

$$P_* = P(W|c_*) = \frac{p(c_1)}{p(c_*)} P_1 + \frac{p(c_2)}{p(c_*)} P_2$$

Ορίζουμε την γενικευμένη Jensen-Shannon απόκλιση μεταξύ κατανομών P_1 και P_2 ως

$$D_{JS}(P_1, P_2) = \frac{p(c_1)}{p(c_*)} D_{KL}(P_1||P_*) + \frac{p(c_2)}{p(c_*)} D_{KL}(P_2||P_*)$$

$D_{KL}(P_1||P_*)$ είναι η KL-απόκλιση μεταξύ των κατανομών P_1 και P_*. Ορίζουμε *C* να είναι το clustering με τα clusters c_1 και c_2, και *C** το clustering αφού συνενώσουμε τα clusters c_1 και c_2 στο cluster c_*.

2. Δείξτε ότι η αύξηση στην εντροπία είναι

$$H(W|C^*) - H(W|C) = p(c_*)\, D_{JS}(P_1, P_2)$$

Υπάρχουν μεθοδολογίες για clustering οι οποίες στοχεύουν στην ελαχιστοποίηση της εντροπίας του clustering, και χρησιμοποιούν την αύξηση της εντροπίας ως μέτρο απόστασης μεταξύ των clusters. Μπορούμε μετά να χρησιμοποιήσουμε ένα agglomerative αλγόριθμο, ή τον *k*-means αλγόριθμο.

Θεωρείστε ένα απλό παράδειγμα όπου το λεξικό αποτελείται από μόνο δύο λέξεις, W = {“sports”, “politics”}, και έχουμε τρία κείμενα d_1, d_2, d_3. Το κείμενο d_1περιέχει δύο εμφανίσεις της λέξης “sports”, το κείμενο d_2 περιέχει τρείς εμφανίσεις της λέξης “sports”, και το κείμενο d_3περιέχει πέντε εμφανίσεις της λέξης “politics”.

3. Ποια είναι η εντροπία του κάθε κειμένου και τι σημαίνει αυτό? Ποια είναι η απόσταση των δύο πιο κοντινών κειμένων και τι σημαίνει αυτό? Ποια είναι η εντροπία ενός cluster που περιέχει και τα τρία κείμενα?

Άσκηση 4

Ο σκοπός αυτής της άσκησης είναι να πειραματιστείτε με το K-means clustering. Μπορείτε να δημιουργήσετε τη δική σας υλοποίηση του K-means αλγορίθμου που περιγράψαμε στην τάξη, ή να χρησιμοποιήσετε κάποια διαθέσιμη υλοποίηση. (Ο K-means υλοποιείται στο MATLAB, και στο WEKA, και υπάρχουν πολλές διαθέσιμες υλοποιήσεις online, αλλά ο αλγόριθμος είναι απλός και ίσως είναι πιο εύκολο να κάνετε τη δική σας υλοποίηση).

Θα εφαρμόσετε τον αλγόριθμο στα Twitter δεδομένα που χρησιμοποιήσατε και για την Πρώτη Σειρά Ασκήσεων. Γι αυτή την σειρά, δεν θα χρησιμοποιήσετε όλα τα δεδομένα. Χρησιμοποιώντας το πεδίο της τοποθεσίας (το 7^{ο} πεδίο στο αρχείο), θα διαλέξετε τους χρήστες που μένουν στις ακόλουθες πόλεις: London, Los Angeles, New York, San Francisco, και Hollywood. Βεβαιωθείτε ότι η μέθοδος με την οποία διαλέγετε τους χρήστες θα βρει αρκετούς από τους διαφορετικούς τρόπους με τους οποίους οι χρήστες δηλώνουν την ίδια πόλη (π.χ., "NYC" για "New York", "Los Angeles, CA" v.s. "Los Angeles"). Επιθεωρήστε τα τελικά δεδομένα και πετάξτε τους χρήστες που έχουν πολλαπλές πόλεις στην τοποθεσία τους.

Αφού μαζέψετε τους παραπάνω χρήστες, πάρετε και το profile description τους και εφαρμόστε την ίδια προεπεξεργασία όπως και στην Πρώτη Σειρά Ασκήσεων (αφαιρέστε stop-words, κλπ). Το αποτέλεσμα θα είναι ένα πραγματικό διάνυσμα για κάθε χρήστη, όπου η κάθε διάσταση αντιστοιχεί σε μία λέξη, και η τιμή του διανύσματος είναι ο αριθμός των εμφανίσεων της λέξης στο description. Η αναπαράσταση του διανύσματος θα εξαρτηθεί από την υλοποίηση του K-means που θα χρησιμοποιήσετε (π.χ., το πιο πιθανό είναι ότι δεν θα χρειαστεί να δημιουργήσετε τις μηδενικές τιμές). Εφαρμόστε τον K-means αλγόριθμο σε αυτά τα διανύσματα, για K = 5.

Αξιολογήστε τα αποτελέσματα του clustering συγκρίνοντας με τις 5 κλάσεις που ορίζονται από τις 5 πόλεις. Αυτή η πληροφορία θα χρησιμοποιηθεί ως μια εξωτερική αξιολόγηση του clustering. Δημιουργείστε τον πίνακα σύγχυσης μεταξύ των clusters και των κλάσεων, και υπολογίστε το Precision και Recall για κάθε cluster.

Παραδώστε τον κώδικα σας, το αρχείο με τους επιλεγμένους χρήστες (δύο στήλες, μία με την τοποθεσία και μία με την περιγραφή) και το ίδιο αρχείο με τους χρήστες ομαδοποιημένους σε clusters. Παραδώστε επίσης μία αναφορά με τις λεπτομέρειες του πειράματος σας: την υλοποίηση που χρησιμοποιήσατε για τον K-means; Την προεπεξεργασία που κάνατε στα δεδομένα και τον τρόπο επιλογής των χρηστών; τον πίνακα σύγχυσης και τις τιμές για τα Precision και Recall.

Όπως και στην Πρώτη Σειρά, αν υπάρχει κάποιο άλλο dataset το οποίο σας ενδιαφέρει να αναλύσετε και να ομαδοποιήσετε, επικοινωνήστε μαζί μου με την πρόταση σας.

Άσκηση 5

Στο μάθημα περιγράψαμε την Minimum Description Length τεχνική, και την εφαρμογή της στο πρόβλημα του co-clustering. Στην άσκηση αυτή θα πρέπει να εφαρμόσετε την τεχνική στο πρόβλημα της κατάτμησης (segmentation) μιας μονοδιάστατης ακολουθίας. Η είσοδος είναι μια ακολουθία από n αριθμούς με τιμές 0 ή 1. Ο στόχος είναι να σπάσουμε την ακολουθία σε τμήματα, έτσι ώστε το συνολικό κόστος της κωδικοποίησης της ακολουθίας να ελαχιστοποιηθεί. Σημειώστε ότι σε αυτή την εκδοχή του προβλήματος, ο αριθμός των τμημάτων K **δεν** μας δίνεται ως είσοδος.

1. Ορίστε το κόστος της κωδικοποίησης μιας κατάτμησης. Ξεχωρίστε το κόστος της κωδικοποίησης του μοντέλου (model cost), και το κόστος της κωδικοποίησης των δεδομένων βάσει του μοντέλου (data cost). **Υπόδειξη**: μία κατάτμηση σε K τμήματα ορίζεται από K-1 συνοριακά σημεία (boundary points).
2. Δώστε ένα ευρηστικό (heuristic) αλγόριθμο για το πρόβλημα της εύρεσης μιας κατάτμησης που ελαχιστοποιεί το κόστος. Περιγράψτε τον αλγόριθμο στα ελληνικά και/η με ψευδοκώδικα.
3. Το πρόβλημα της κατάτμησης μπορεί να λυθεί βέλτιστα σε πολυωνυμικό χρόνο χρησιμοποιώντας δυναμικό προγραμματισμό. Ορίστε την αναδρομική σχέση του δυναμικού προγραμματισμού που λύνει το πρόβλημα, και τον πίνακα του δυναμικού προγραμματισμού.
4. **Bonus**: Υλοποιείστε ένα αλγόριθμο που να λύνει το πρόβλημα. Τεστάρετε τον αλγόριθμο σας σε μια ακολουθία 1000 τιμών την οποία θα δημιουργήσετε ως εξής:
    a. Διαλέξτε τυχαία 2 σημεία διάσπασης.
    b. Δημιουργείστε τυχαίες τιμές 0/1 για κάθε τμήμα, με διαφορετικές πιθανότητες: στο πρώτο τμήμα η πιθανότητα της τιμής 1 είναι 0.7, στο δεύτερο 0.3, και στο τρίτο 0.9.
    c. Παραδώσετε ένα αρχείο με τα πραγματικά σημεία διάσπασης και αυτά που βρίσκει ο αλγόριθμος σας για 3 διαφορετικά τυχαία inputs.