Υπολογιστική Εργασία Αναγνώρισης Προτύπων
Ακαδημαϊκό Έτος 2022 – 2023
Γ. Τσιχριντζής Δ. Σωτηρόπουλος

Θέμα: Αλγόριθμοι Σύστασης με Χρήση Τεχνητών Νευρωνικών Δικτύων και Τεχνικών Ομαδοποίησης Δεδομένων

Η συγκεκριμένη εργασία έχει ως στόχο την ανάπτυξη αλγορίθμων σύστασης ταινιών κάνοντας χρήση τεχνητών νευρωνικών δικτύων και τεχνικών συσταδοποίησης. Το σύνολο των δεδομένων επάνω στο οποίο θα εργαστείτε μπορείτε να το κατεβάσετε από την διεύθυνση https://ieee-dataport.org/open-access/imdb-movie-reviews-dataset και αφορά σε ένα σύνολο χρηστών $\boldsymbol{U} = \{U_1, U_2, \cdots, U_N, \}$ οι οποίοι εκφράζουν προτιμήσεις επάνω σε ένα σύνολο αντικειμένων-ταινιών $\boldsymbol{I} = \{I_1, I_2, \cdots, I_N\}$. Ο χρήστης $u \in \boldsymbol{U}$ αποδίδει στην ταινία $i \in \boldsymbol{I}$ τον βαθμό προτίμησης $R(u,i) \in \boldsymbol{R_o} = \{1,2,3,4,5,6,7,8,9,10\} \cup \{0\}$, όπου η τιμή 0 υποδηλώνει την απουσία αξιολόγησης και όχι μηδενικό βαθμό προτίμησης.

Το σύνολο $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$ των διαθέσιμων δεδομένων αποτελείται από εγγραφές της μορφής $\boldsymbol{x}_k = [u_k, i_k, r_k, t_k]$ όπου $u_\kappa \in \boldsymbol{U}$ και $i_\kappa \in \boldsymbol{I}$ με $r_\kappa = R(u_\kappa, i_\kappa) \in \boldsymbol{R_o}$. Η τιμή t_k αντιστοιχεί στην ημερομηνία καταχώρησης της συγκεκριμένης εγγραφής.

Προ-επεξεργασία Δεδομένων:

1. Να βρείτε το σύνολο των μοναδικών χρηστών $\boldsymbol{U}$ και το σύνολο των μοναδικών αντικειμένων $\boldsymbol{I}$.
2. Θεωρείστε την συνάρτηση $\Phi: \mathbf{U} \to \boldsymbol{P}(\boldsymbol{I})$ (όπου $\boldsymbol{P}(\mathbf{I})$ το δυναμοσύνολο του $\mathbf{I}$) η οποία $\forall\, u \in \mathbf{U}$ επιστρέφει το σύνολο $\varphi(u) \subset \boldsymbol{I}$ των αντικειμένων που αξιολογήθηκαν από τον χρήστη u. Μπορούμε να γράψουμε, επομένως, ότι $\varphi(u) = \{i \in \boldsymbol{I}: R(u,i) > 0\}$. Να περιορίσετε τα σύνολα των μοναδικών χρηστών $\mathbf{U}$ και μοναδικών αντικειμένων $\boldsymbol{I}$ στα αντίστοιχα σύνολα $\widehat{\mathbf{U}}$ και $\widehat{\mathbf{I}}$ έτσι ώστε $\forall\, u \in \widehat{\boldsymbol{U}}, R_{min} \leq |\varphi(u)| \leq R_{max}$ όπου R_{min} και R_{max} ο ελάχιστος απαιτούμενος και ο μέγιστος επιτρεπτός αριθμός αξιολογήσεων ανά χρήστη. Θεωρήστε προφανώς ότι $\left|\widehat{\boldsymbol{U}}\right| = \boldsymbol{n} < \boldsymbol{N}$ και $\left|\widehat{\boldsymbol{I}}\right| = \boldsymbol{m} < \boldsymbol{M}$.
3. Να δημιουργήσετε και να αναπαραστήσετε γραφικά τα ιστογράμματα συχνοτήτων για το πλήθος αλλά και για το χρονικό εύρος των αξιολογήσεων του κάθε χρήστη.
4. Δημιουργήστε μια εναλλακτική αναπαράσταση του συνόλου των δεδομένων ως ένα σύνολο διανυσμάτων προτιμήσεων $\boldsymbol{R} = \{\boldsymbol{R_1}, \boldsymbol{R_2}, \cdots, \boldsymbol{R_n}\}$ με $\boldsymbol{R_j} = \boldsymbol{R}(u_j) \in \boldsymbol{R_o}^{\,\boldsymbol{m}}, \forall\, j \in [n]$. Συγκεκριμένα, μπορούμε να γράψουμε ότι:

$$\boldsymbol{R_j}(k) = \begin{cases} R(u_j, i_k), R(u_j, i_k) > 0; \\ 0, \qquad\quad R(u_j, i_k) \;>\; 0. \end{cases} \quad (\mathbf{1})$$

Αλγόριθμοι Ομαδοποίησης Δεδομένων

1. Να οργανώσετε το περιορισμένο σύνολο των χρηστών $\widehat{\mathbf{U}}$ σε L συστάδες (clusters) της μορφής $\widehat{\mathbf{U}} = \boldsymbol{U_{G_1}} \cup \boldsymbol{U_{G_2}} \cup \cdots \cup \boldsymbol{U_{G_L}}$ έτσι ώστε $\boldsymbol{U_{G_a}} \cap \boldsymbol{U_{G_b}} \neq \emptyset, \forall \boldsymbol{a} \neq \boldsymbol{b}$ βασιζόμενοι στην διανυσματική αναπαράσταση των προτιμήσεών τους μέσω του συνόλου $\boldsymbol{R}$. Η διαδικασία ομαδοποίησης των διανυσμάτων προτιμήσεων των χρηστών $\boldsymbol{R_j} = \boldsymbol{R}(u_j) \in \boldsymbol{R_o}^{\,\boldsymbol{m}}, \forall\, j \in [n]$ επάνω στο περιορισμένο σύνολο των αντικειμένων $\widehat{\mathbf{I}}$, μπορεί να πραγματοποιηθεί με κατάλληλη παραμετροποίηση του αλγορίθμου **k-means** μεταβάλλοντας την μετρική που

αποτιμά την απόσταση μεταξύ δύο διανυσμάτων προτιμήσεων για ένα ζεύγος χρηστών u και v ως $\boldsymbol{dist}(\boldsymbol{R_u}, \boldsymbol{R_v})$. Συγκεκριμένα, μπορείτε να χρησιμοποιήσετε τις παρακάτω μετρικές (*):

a. $\boldsymbol{dist_{euclidean}}(\boldsymbol{R_u}, \boldsymbol{R_v}) = \sqrt{\sum_{k=1}^{m} |\boldsymbol{R_u}(k) - \boldsymbol{R_v}(k)|^2 \, \lambda_u(k)\lambda_v(k)}$ $(\mathbf{2})$ έτσι ώστε:

$$\lambda_j(k) = \begin{cases} 1, & R(u_j, i_k) > 0; \\ 0, & R(u_j, i_k) > 0. \end{cases} \quad (\mathbf{3})$$

Η συνάρτηση $\lambda_j()$ αποτιμά το αν ο χρήστης u_j έχει αξιολογήσει ή όχι το αντικείμενο i_k. Προφανώς, μέσω της συγκεκριμένης συνάρτηση μπορούμε αν υπολογίσουμε την τιμή του $|\varphi(u)|$ που αποτιμά το πλήθος των αξιολογήσεων που παρείχε συνολικά ο χρήστης u ως: $|\varphi(u)| = \sum_{k=1}^{n} \lambda_u(k)$.

b. $$\boldsymbol{dist_{cosine}}(\boldsymbol{R_u}, \boldsymbol{R_v}) = 1 - \left| \frac{\sum_{k=1}^{m} \boldsymbol{R_u}(k)\boldsymbol{R_v}(k)\lambda_u(k)\lambda_v(k)}{\sqrt{\sum_{k=1}^{m} \boldsymbol{R_u}(k)^2\lambda_u(k)\lambda_v(k)}\sqrt{\sum_{k=1}^{m} \boldsymbol{R_v}(k)^2\lambda_u(k)\lambda_v(k)}} \right| \quad (\mathbf{4})$$

c. Να αναπαραστήσετε γραφικά τις συστάδες των χρηστών που αναγνωρίστηκαν από τον αλγόριθμο k-means για κάθε μία από τις παραπάνω μετρικές για διάφορες τιμές της παραμέτρου L.

d. Να σχολιάσετε την αποτελεσματικότητα των συγκεκριμένων μετρικών στην αποτίμηση της ομοιότητας μεταξύ ενός ζεύγους διανυσμάτων προτιμήσεων χρηστών R_u και R_v.

*** Οι μετρικές που αναπαρίστανται στις εξισώσεις (2) και (4) υπολογίζονται για κάθε ζεύγος διανυσμάτων $\boldsymbol{R_u}$ και $\boldsymbol{R_v}$ μόνο για το υποσύνολο των αντικειμένων που έχουν αξιολογηθεί και από τους δύο χρήστες $\boldsymbol{u}$ και $\boldsymbol{v}$ μέσω της βοηθητικής συνάρτησης $\boldsymbol{\lambda}()$.**

Αλγόριθμοι Παραγωγής Συστάσεων με Χρήση Τεχνητών Νευρωνικών Δικτύων

2. Να δημιουργήσετε μια εναλλακτική οργάνωση του περιορισμένου συνόλου των χρηστών σε L συστάδες $\widehat{\mathbf{U}} = \boldsymbol{U_{G_1}} \cup \boldsymbol{U_{G_2}} \cup \cdots \cup \boldsymbol{U_{G_L}}$ έτσι ώστε $\boldsymbol{U_{G_a}} \cap \boldsymbol{U_{G_b}} \neq \emptyset, \forall \boldsymbol{a}, \boldsymbol{b} \in [\boldsymbol{L}]$ με $\boldsymbol{a} \neq \boldsymbol{b}$ κάνοντας χρήση της παρακάτω μετρικής (**):

$$dist(u, v) = 1 - \frac{|\varphi(u) \cap \varphi(v)|}{|\varphi(u) \cup \varphi(v)|} \quad (\mathbf{5})$$

**** Ο αλγόριθμος ομαδοποίησης για την αντιμετώπιση του συγκεκριμένου ερωτήματος είναι της επιλογής σας. Βασική προϋπόθεση, ωστόσο, είναι να μπορεί να λειτουργήσει επάνω στον τετραγωνικό πίνακα των αποστάσεων μεταξύ των χρηστών που περιγράφει η σχέση (5).**

a) Να εξηγήσετε τι εκφράζει η συγκεκριμένη μετρική και να προσδιορίσετε τα μειονεκτήματά της σε σχέση με τις μετρικές που περιγράφονται στις σχέσεις (2) και (4) υπό το πρίσμα της ιδιαίτερης οργάνωσης των χρηστών που μπορεί να επιφέρει.

b) Η μετρική που περιγράφεται μέσω της σχέσης (5) μπορεί να χρησιμοποιηθεί προκειμένου να προσδιοριστεί το σύνολο $N_k(u_a) = \{u_a^{(1)}, u_a^{(2)}, \cdots, u_a^{(k)}\}$ των k πλησιέστερων γειτόνων ενός χρήστη $u_a \in \boldsymbol{U_{G_a}}, \forall \boldsymbol{a} \in [\boldsymbol{L}]$. Επομένως, για τον εκάστοτε χρήστη u_a της εκάστοτε συστάδας $\boldsymbol{U_{G_a}}$ μπορούμε να το διάνυσμα τον προσωπικών του προτιμήσεων R_{u_a} καθώς και τα διανύσματα των k πλησιέστερων γειτόνων του $R_{u_a^{(1)}}$, $R_{u_a^{(2)}}$,..., $R_{u_a^{(\kappa)}}$ εντός της συστάδας $\boldsymbol{U_{G_a}}$. Στόχος του συγκεκριμένου ερωτήματος είναι να αναπτύξετε ένα πολυστρωματικό νευρωνικό δίκτυο για κάθε συστάδα χρηστών $\boldsymbol{U_{G_a}}$ $\mu\varepsilon$ $a \in [L]$ το οποίο θα προσεγγίζει τις αξιολογήσεις του κάθε χρήστη εντός αυτής μέσω των αξιολογήσεων των k πλησιέστερων γειτόνων του μέσω μιας συνάρτησης της μορφής:

$$R_{u_a} = f_a\left(R_{u_a^{(1)}}, R_{u_a^{(2)}}, \cdots R_{u_a^{(k)}}\right), \forall u_a \in \boldsymbol{U_{G_a}}, \forall a \in [L] (\boldsymbol{6})$$

c) Το σύνολο των χρηστών της κάθε συστάδας μπορεί να διαμεριστεί περεταίρω σε ένα υποσύνολο χρηστών για εκπαίδευση του νευρωνικού δικτύου $\boldsymbol{U_{G_a}^{train}}$ και σε ένα υποσύνολο χρηστών για τον έλεγχο της επίδοσης του νευρωνικού δικτύου $\boldsymbol{U_{G_a}^{test}}$ έτσι ώστε $\boldsymbol{U_{G_a}} = \boldsymbol{U_{G_a}^{train}} \cup \boldsymbol{U_{G_a}^{test}}$ με $\boldsymbol{U_{G_a}^{train}} \cap \boldsymbol{U_{G_a}^{test}} = \emptyset$. Η καταλληλότερη διαμόρφωση των δεδομένων για την εκπαίδευση των νευρωνικών δικτύων αυτού του ερωτήματος θα μπορούσε να αναπαρασταθεί ως εξής:

Έστω ότι το σύνολο των χρηστών που μετέχουν στην συστάδα $\boldsymbol{U_{G_a}}$ δίνεται από την σχέση: $\boldsymbol{U_{G_a}} = \{u_{a,1}, u_{a,2}, \cdots, u_{a,n_a}\}$ με $n_a = \left|\boldsymbol{U_{G_a}}\right|$. Τότε το σύνολο των διανυσμάτων χαρακτηριστικών και το αντίστοιχο σύνολο των ετικετών θα μπορούσε να οργανωθεί σε έναν πίνακα της μορφής:

$$\begin{bmatrix} R_{u_{a,1}^{(1)}} & R_{u_{a,1}^{(2)}} & \cdots & R_{u_{a,1}^{(k)}} \\ R_{u_{a,2}^{(1)}} & R_{u_{a,2}^{(2)}} & \cdots & R_{u_{a,2}^{(k)}} \\ \vdots & \vdots & \vdots & \vdots \\ R_{u_{a,n_a}^{(1)}} & R_{u_{a,n_a}^{(2)}} & \cdots & R_{u_{a,n_a}^{(k)}} \end{bmatrix} \text{ και } \begin{bmatrix} R_{u_{a,1}} \\ R_{u_{a,2}} \\ \vdots \\ R_{u_{a,n_a}} \end{bmatrix}$$

d) Η προσεγγιστική ακρίβεια των παραπάνω νευρωνικών δικτύων μπορεί να μετρηθεί μέσω της μετρικής του μέσου απόλυτου σφάλματος ανάμεσα στις πραγματικές και τις εκτιμώμενες αξιολογήσεις των χρηστών. Να παρουσιάσετε πίνακες των αποτελεσμάτων σας τόσο για την ακρίβεια εκπαίδευσης όσο και για την ακρίβεια ελέγχου για κάθε συστάδα χρηστών.

Μπορείτε να εργαστείτε σε ομάδες 3 φοιτητών το πολύ. Η υλοποίηση της εργασίας μπορεί να πραγματοποιηθεί σε MATLAB ή Python. Ο κώδικας της εργασίας σας θα πρέπει να συνοδεύεται από κείμενο αναλυτικής τεκμηρίωσης.

Καλή Επιτυχία!