ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Διδάσκοντες: Σ. Λυκοθανάσης, Δ. Κουτσομητρόπουλος

Ακαδημαϊκό Έτος 2019-2020

Εργαστηριακή Άσκηση
Μέρος Β΄

Β. Υλοποίηση Γενετικού Αλγορίθμου για Συνεργατικό Φιλτράρισμα (Σύσταση Ταινιών)

Στην εργασία αυτή σας ζητείται να προτείνετε και να υλοποιήσετε **Γενετικό Αλγόριθμο** που θα χρησιμοποιηθεί για τη εξατομικευμένη σύσταση ταινιών σε χρήστες. Προς το σκοπό αυτό θα αξιοποιηθεί η μέθοδος του *συνεργατικού φιλτραρίσματος* (collaborative filtering -CF), όπως και στο μέρος Α.
Σκοπός του αλγορίθμου είναι να υπολογίσει το βέλτιστο διάνυσμα που αντιστοιχεί στις αξιολογήσεις ενός συγκεκριμένου χρήστη της επιλογής σας για όλες τις ταινίες που περιέχονται στο dataset, συμπεριλαμβανομένων και αυτών που δεν έχει ακόμα δει.
Παρακάτω φαίνεται ένα δείγμα αξιολογήσεων που περιέχονται στο σύνολο δεδομένων movieLens 100K (https://grouplens.org/datasets/movielens/100k/):

user_id	item_id	rating	timestamp
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596

Η βασική ιδέα για την αξιοποίηση του ΓΑ για συστάσεις ταινιών είναι η εξής: Αρχικά επιλέγεται ένας συγκεκριμένος χρήστης. Στη συνέχεια, εντοπίζονται οι χρήστες εκείνοι που μοιάζουν περισσότερο στον συγκεκριμένο με βάση τις υπάρχουσες αξιολογήσεις. Στόχος του αλγορίθμου είναι να βρει τις τιμές εκείνες για τις αξιολογήσεις που λείπουν, ώστε ο συγκεκριμένος χρήστης να μοιάζει όσο το δυνατόν περισσότερο με τη γειτονιά του.

Β1. Σχεδιασμός ΓΑ *[30 μονάδες]*

α) Κωδικοποίηση: Να προτείνετε μια κωδικοποίηση για τα άτομα του πληθυσμού. Λάβετε υπόψη τα παρακάτω:

i. Ένας χρήστης αναπαρίσταται ως ένα αραιό διάνυσμα *M* διαστάσεων, όσες και οι ταινίες, με τιμές τις αξιολογήσεις του χρήστη σε κάθε θέση, π.χ. u^T = [1, 2, - , 5, - , ….], όπου – δηλώνει την έλλειψη αξιολόγησης.
ii. Οι αξιολογήσεις, όπου υπάρχουν, έχουν ακέραιες τιμές στο διάστημα [1,5].
iii. Τα άτομα του πληθυσμού είναι πιθανά διανύσματα αξιολογήσεων που αντιστοιχούν στον τρέχοντα χρήστη.

β) Πλεονάζουσες τιμές: Ανάλογα με την κωδικοποίηση που εφαρμόσατε στο (α) είναι πιθανό να προκύψουν πλεονάζουσες τιμές, για παράδειγμα, τιμές > 5 λόγω δυαδικής κωδικοποίησης. Περιγράψτε πώς θα αντιμετωπίσετε το πρόβλημα αυτό. Εξετάστε αν μπορείτε να αποφύγετε τις πλεονάζουσες τιμές, αναθεωρώντας την κωδικοποίηση που προτείνατε στο (α).

γ) Αρχικός πληθυσμός: Περιγράψτε μια διαδικασία για τη δημιουργία αρχικού πληθυσμού ατόμων. Τα άτομα του πληθυσμού είναι πιθανά διανύσματα αξιολογήσεων που αντιστοιχούν στον τρέχοντα χρήστη, με τιμές ιδίως για τις ταινίες που δεν έχει δει (-).

i. *Τιμές για τις ταινίες που **δεν** έχει δει:* Τι είδους αρχικές τιμές θα επιλέξετε για τις ταινίες που δεν έχει δει ο χρήστης; Μπορεί να αντιστοιχούν στο μέσο όρο, όπως στο μέρος Α;
ii. *Τιμές για τις ταινίες που έχει ήδη δει:* Τι θα συμβεί με τις τιμές για τις ταινίες που ο χρήστης έχει δει; Θα τις συμπεριλάβετε στα γονίδια των ατόμων του πληθυσμού; Θα πρέπει να παραμένουν σταθερές; Σημειώστε ότι μια λύση στην οποία οι τιμές για τις αξιολογήσεις ταινιών που έχει δει ο χρήστης αποκλίνουν από τις πραγματικές, μπορεί να καταστήσει τη λύση μη νόμιμη.

δ) Διαδικασία επιδιόρθωσης: Εφόσον στα γονίδια κάθε ατόμου περιλαμβάνονται και τιμές για τις ταινίες που ήδη έχει αξιολογήσει ο χρήστης, είναι πιθανό να προκύψουν λύσεις που αποκλίνουν από αυτές τις τιμές. Αν η κωδικοποίησή σας είναι πιθανό να δημιουργεί μη νόμιμες λύσεις, προδιαγράψτε μια διαδικασία χειρισμού των μη νόμιμων λύσεων, σχολιάζοντας και αξιολογώντας τις παρακάτω εναλλακτικές:

i. Απόρριψη της μη νόμιμης λύσης από τον πληθυσμό και αντικατάστασής της από κάποιο άλλο άτομο (τυχαία ή με ελιτισμό).
ii. Επιδιόρθωση: Διαδικασία επιδιόρθωσης (repair procedure) η οποία αντιστοιχίζει τη μη νόμιμη λύση σε μια νόμιμη, π.χ. αντικατάσταση κάθε φορά των γνωστών αξιολογήσεων με τις πραγματικές τιμές.
iii. Εφαρμογή ποινής: Μια μη νόμιμη λύση γίνεται αποδεκτή, αλλά της εφαρμόζεται ανάλογη ποινή από την συνάρτηση καταλληλότητας. Να περιγράψετε μια διαδικασία εφαρμογής ποινής για τέτοιες λύσεις.

Αξιολογήστε τις παραπάνω μεθόδους και προτείνετε την καταλληλότερη για το πρόβλημά σας.

ε) Εύρεση γειτονιάς χρήστη: Υπολογίστε την απόσταση του χρήστη από όλους τους υπόλοιπους στο σύνολο εκπαίδευσης (βλ. παρακάτω Β4) και βρείτε τους top-10 που είναι πιο κοντά (*γειτονιά του χρήστη*). Για τον υπολογισμό της απόστασης μπορούν να χρησιμοποιηθούν διάφορες μετρικές, όπως *ευκλείδεια απόσταση*, *απόσταση Manhattan*, *συνημίτονο* και *συσχέτιση Pearson*. Να χρησιμοποιήσετε την μετρική Pearson και να σχολιάσετε την καταλληλότητά της, σε σχέση και με τις υπόλοιπες, για την αποφυγή της πόλωσης στις βαθμολογίες. Επίσης να αιτιολογήσετε πώς θα αντιμετωπίσετε τις ελλιπείς τιμές, ειδικά για τον υπολογισμό της γειτονιάς (Θα τις αγνοήσετε; Θα θεωρήσετε 0; Τυχαίες τιμές; Μέσο όρο;).

στ) Συνάρτηση καταλληλότητας: Ένα άτομο είναι πιο κατάλληλο από άλλα, εφόσον βρίσκεται πιο κοντά στη γειτονιά του χρήστη. Επομένως ως βάση για τη συνάρτηση καταλληλότητας μπορεί να χρησιμοποιηθεί η μεγιστοποίηση του συντελεστή ομοιότητας Pearson του διανύσματος του χρήστη με τους χρήστες που βρίσκονται στη γειτονιά του. Μπορείτε να χρησιμοποιήσετε είτε άθροισμα είτε Μ.Ο. Επειδή η τιμή του συντελεστή Pearson κινείται στο [-1, 1], εξετάστε αν χρειάζεται να κλιμακώστε κατάλληλα τη συνάρτηση καταλληλότητας, ώστε να παίρνει μόνο μη αρνητικές τιμές. Ποια θα είναι η μέγιστη τιμή που μπορεί να έχει;

ζ) Γενετικοί Τελεστές: Με βάση την κωδικοποίηση που επιλέξατε να προτείνετε τους τελεστές επιλογής, διασταύρωσης και μετάλλαξης που θα χρησιμοποιήσετε.

i. Ειδικά για την επιλογή, να αξιολογήσετε τη χρήση *ρουλέτας με βάση το κόστος, με βάση την κατάταξη* και *τουρνουά.*
ii. Ειδικά για τη διασταύρωση, να αξιολογήσετε την καταλληλότητα των ακόλουθων τελεστών: *Διασταύρωση μονού σημείου, διασταύρωση πολλαπλού σημείου, ομοιόμορφη διασταύρωση, OX* και *PMX*.
iii. Ειδικά για τη μετάλλαξη, να αξιολογήσετε τη χρήση *ελιτισμού*.

Β2. Υλοποίηση ΓΑ *[30 μονάδες]*
Να γράψετε ένα πρόγραμμα, σε οποιοδήποτε περιβάλλον ή γλώσσα προγραμματισμού, που να υλοποιεί τον γενετικό αλγόριθμο που σχεδιάσετε.

Β3. Αξιολόγηση και Επίδραση Παραμέτρων *[20 μονάδες]*
α) Να τρέξετε τον αλγόριθμο για τις τιμές των παραμέτρων που φαίνονται στον παρακάτω πίνακα και να τον συμπληρώσετε. Ο αλγόριθμος θα τερματίζει όταν πληρούνται ένα ή περισσότερα από τα *κριτήρια τερματισμού*, δηλαδή όταν:

i. το καλύτερο άτομο της κάθε γενιάς πάψει να βελτιώνεται για ορισμένο αριθμό γενεών ή
ii. βελτιώνεται κάτω από ένα ποσοστό (<1%) ή
iii. έχει ξεπεραστεί ένας προκαθορισμένος αριθμός γενεών (π.χ. 1000)

Α/Α	ΜΕΓΕΘΟΣ ΠΛΗΘΥΣΜΟΥ	ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩΣΗΣ	ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ	ΜΕΣΗ ΤΙΜΗ ΒΕΛΤΙΣΤΟΥ	ΜΕΣΟΣ ΑΡΙΘΜΟΣ ΓΕΝΕΩΝ
1	20	0.6	0.00		
2	20	0.6	0.01		
3	20	0.6	0.10		
4	20	0.9	0.01		
5	20	0.1	0.01		
6	200	0.6	0.00		
7	200	0.6	0.01		
8	200	0.1	0.01		
9	200	0.9	0.01		

Προσοχή: *Επειδή οι ΓΑ είναι στοχαστικοί αλγόριθμοι και συνεπώς δεν εξασφαλίζουν την ίδια απόδοση σε κάθε εκτέλεσή τους, θα πρέπει να εκτελέσετε τον αλγόριθμο τουλάχιστον δέκα φορές για κάθε περίπτωση. Στον πίνακα να σημειώσετε το μέσο όρο της απόδοσης της καλύτερης λύσης σε κάθε τρέξιμο.*

β) Για κάθε περίπτωση του παραπάνω πίνακα να σχεδιάστε την καμπύλη εξέλιξης (απόδοση/αριθμό γενιών) της καλύτερης λύσης (της μέσης τιμής αυτής, σε κάθε τρέξιμο).

γ) Με βάση αυτές τις καμπύλες, αλλά και τα αποτελέσματα του παραπάνω πίνακα, να διατυπώσετε αναλυτικά τα συμπεράσματά σας σχετικά με την επίδραση της κάθε παραμέτρου (μέγεθος πληθυσμού, πιθανότητα διασταύρωσης, πιθανότητα μετάλλαξης) στη σύγκλιση του αλγορίθμου.

Β4. Αξιολόγηση Συστάσεων *[20 μονάδες]*
α) Εφαρμόστε holdout για να δημιουργήσετε σύνολο εκπαίδευσης και σύνολο ελέγχου. Για κάθε χρήστη κρατήστε χωριστά 10 αξιολογήσεις (~10%). Παρατηρήστε ότι ένα τέτοιο σύνολο ελέγχου ήδη περιλαμβάνεται στο dataset (αρχεία ua.base και ua.test ή ub.base και ub.test). Εναλλακτικά, μπορείτε να χρησιμοποιήσετε 5-fold cross-validation όπως στο μέρος Α και να αναφέρετε μέσους όρους για τις τιμές που ζητούνται.

β) Με τις βέλτιστες τιμές των παραμέτρων που βρήκατε στο Β3, να αναφέρετε RMSE και MAE για τις 10 αξιολογήσεις που περιέχονται στο σύνολο ελέγχου, για έναν χρήστη της επιλογής σας, δίνοντας τις αντίστοιχες γραφικές παραστάσεις ανά γενιά, μαζί με την γραφική παράσταση της καταλληλότητας του βέλτιστου ατόμου ανά γενιά (10 εκτελέσεις λόγω στοχαστικής φύσης).

γ) Να αναφέρετε RMSE και MAE για τουλάχιστον 50 χρήστες (δηλαδή για 50*10=500 αξιολογήσεις), επαναλαμβάνοντας την παραπάνω διαδικασία.

B5. Εφαρμογή ΓΑ για όλους τους χρήστες *[προαιρετικό ερώτημα - 10 μονάδες bonus]*. Θεωρήστε τα άτομα του πληθυσμού, όχι ως διανύσματα για έναν χρήστη, αλλά ως *NxM* αραιούς *πίνακες* όλων των χρηστών. Ο πίνακας θα περιέχει τιμές που αντιστοιχούν στις προβλέψεις αξιολόγησης των χρηστών για τις ταινίες που δεν έχουν δει. Ως αποτέλεσμα, με τον τερματισμό του ΓΑ θα έχει βρεθεί η βέλτιστη λύση για όλους τους χρήστες και όχι μόνο για έναν. Να διατυπώσετε τους γενετικούς τελεστές και να επαναλάβετε τα ζητούμενα Β2, Β3 και Β4.

Παραδοτέα

Η αναφορά που θα παραδώσετε θα πρέπει να περιέχει εκτενή σχολιασμό των πειραμάτων σας καθώς και πλήρη καταγραφή των αποτελεσμάτων και των συμπερασμάτων σας, ανά υπο-ερώτημα. Επίσης, πρέπει να συμπεριλάβετε στην αρχή της αναφοράς σας ένα σύνδεσμο προς τον κώδικα που έχετε χρησιμοποιήσει (σε κάποια file sharing υπηρεσία ή code repo), ώστε να ληφθεί υπόψη.
Μην ξεχάσετε να συμπληρώσετε τα στοιχεία σας, στην αρχή της 1ης σελίδας.

Αξιολόγηση

Η απάντηση των ερωτημάτων Α και Β, έχει βαρύτητα 20% στον τελικό βαθμό του μαθήματος (το σύνολο και των δύο μερών της εργασίας έχει βαρύτητα 40%). Ο βαθμός του Bonus (10%) προστίθεται στο παραπάνω ποσοστό 40%.

Παρατηρήσεις

1. Η αναφορά, σε ηλεκτρονική μορφή, πρέπει να αναρτηθεί στο e-class μέχρι τη Δευτέρα, 8/6/2020, στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να επικοινωνείτε με τον κ. Ανδρέα Ανδριόπουλο, a.andriopoulos@upatras.gr