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 Α΄

Α. Συνεργατικό Φιλτράρισμα με Χρήση Νευρωνικών Δικτύων για Συστάσεις Ταινιών

Το συνεργατικό φιλτράρισμα (collaborative filtering) είναι μια διαδεδομένη τεχνική για συστήματα συστάσεων (recommendation systems). Βασίζεται στην υπόθεση ότι χρήστες με παρόμοια ενδιαφέροντα και αξιολογήσεις για κάποια αντικείμενα θα έχουν παρόμοιες αξιολογήσεις και για άλλα αντικείμενα που ακόμα δεν έχουν αξιολογήσεις. Για παράδειγμα, έστω τρεις χρήστες u1, u2, u3 και οι αξιολογήσεις τους για δύο ταινίες m1, m2 σε κλίμακα 1-5:

	m1	**m2**
u1	5	-
u2	2	4
u3	5	3

Ποια είναι η πιθανότερη αξιολόγηση του χρήστη u1 για την ταινία m2 που θα μπορούσε να προβλεφθεί; Παρατηρούμε ότι, με βάση τις δηλωμένες αξιολογήσεις, ο χρήστης u1 είναι πιο κοντά στο χρήστη u3 από ότι στον u2 και με βάση το συνεργατικό φιλτράρισμα είναι πιθανό η ζητούμενη αξιολόγηση να είναι κοντά στο 3.

Διάφορες τεχνικές έχουν προταθεί κατά καιρούς στη βιβλιογραφία, μεταξύ άλλων και η χρήση μοντέλων μηχανικής μάθησης για την εύρεση αυτών των προβλέψεων.[1] Επομένως στην παρούσα εργασία θα εξετάσετε τη χρήση ενός πολυεπίπεδου ΤΝΔ για την πρόβλεψη αξιολογήσεων χρηστών σε ταινίες με εφαρμογή συνεργατικού φιλτραρίσματος. Για το σκοπό αυτό θα αξιοποιηθεί το σύνολο δεδομένων *MovieLens 100K* (https://grouplens.org/datasets/movielens/100k/) που περιλαμβάνει ~100K αξιολογήσεις από N = 943 χρήστες για M = 1682 ταινίες. Τα δεδομένα των αξιολογήσεων περιέχονται στο αρχείο u.data, το οποίο περιέχει τις παρακάτω στήλες, χωρισμένες με tab:

user id | item id | rating | timestamp

Για την υλοποίηση των αλγορίθμων μπορείτε να χρησιμοποιήσετε οποιοδήποτε περιβάλλον, βιβλιοθήκη ή γλώσσα προγραμματισμού κρίνετε σκόπιμο. Ενδεικτικά αναφέρονται *MatLab, WEKA, Azure ML Studio, Google Colaboratory, TensorFlow, SciKit-Learn.*

[1] Απουσία χαρακτηριστικών (features) για τους χρήστες και τις ταινίες, η ιδέα είναι ότι χρήστες και ταινίες αναπαρίστανται από λανθάνοντα διανύσματα χαρακτηριστικών που καλούνται *ενσωματώσεις* (embeddings), τα οποία δε γνωρίζουμε ακόμα και που το ΤΝΔ καλείται να «μάθει». Ουσιαστικά, οι τιμές των διανυσμάτων θα αποθηκευτούν στα βάρη του δικτύου. Η ζητούμενη αξιολόγηση θα είναι το εσωτ. γινόμενο των δύο διανυσμάτων χρήστη-ταινίας. Έστω U ο πίνακας διάστασης NxK που προκύπτει από τις ενσωματώσεις των χρηστών και F ο πίνακας διάστασης KxM που προκύπτει από τις ενσωματώσεις των ταινιών. Θα πρέπει το γινόμενο UF να είναι ο πίνακας των αξιολογήσεων όλων των χρηστών για όλες τις ταινίες. Με τον τρόπο αυτό επιτυγχάνεται μείωση της διασταστικότητας, αφού αντί για έναν πίνακα NxM, χρειάζεται να μάθουμε δύο πίνακες μικρότερης διάστασης. Οι τιμές των ζητούμενων πινάκων θα αποθηκεύονται στα βάρη του δικτύου. Προφανώς θα πρέπει K<<M,N

Το ζητούμενο στην εργασία αυτή είναι να κατασκευαστεί και να εκπαιδευτεί ένα ΤΝΔ που θα δέχεται στην είσοδο ένα χρήστη και θα παράγει στην έξοδο τις αξιολογήσεις για όλες τις ταινίες του συνόλου δεδομένων.

Α1. Προεπεξεργασία και Προετοιμασία δεδομένων [*20 μονάδες*]

Προσοχή: Ό,τι μετασχηματισμοί εφαρμοστούν στα δεδομένα του συνόλου εκπαίδευσης, οι ίδιοι θα πρέπει να εφαρμοστούν και στα δεδομένα του συνόλου ελέγχου ή εναλλακτικά να αντιστραφούν πρώτου μετρηθούν οι μετρικές αξιολόγησης παρακάτω.

α) *Κεντράρισμα (centering):* Οι τιμές των αξιολογήσεων είναι ακέραιες και κινούνται στο διάστημα [1,5]. Ωστόσο, πολλές αξιολογήσεις χρηστών μπορεί να εμφανίζουν κάποιο βαθμό *πόλωσης*. Για παράδειγμα, υπάρχουν χρήστες που βαθμολογούν συστηματικά με πολύ χαμηλούς βαθμούς (1-3) και άλλοι χρήστες με υψηλούς βαθμούς (3-5), ενώ η σχετική κατάταξη των ταινιών παραμένει η ίδια. Μια μέθοδος για να μετριαστεί αυτό το φαινόμενο καλείται *κεντράρισμα* (centering) των δεδομένων και επιτυγχάνεται αφαιρώντας το μέσο όρο των αξιολογήσεων ενός χρήστη από όλες τις βαθμολογίες που έχει δώσει.
Εξετάστε τη χρησιμότητα της μεθόδου αυτής για το συγκεκριμένο πρόβλημα και εφαρμόστε τη στα δεδομένα εκπαίδευσης, αν κρίνετε σκόπιμο.
Ποιο είναι το νέο διάστημα στο οποίο θα κινούνται οι τιμές, εφόσον εφαρμοστεί κεντράρισμα;

β) *Ελλιπείς τιμές:* Δεν υπάρχουν αξιολογήσεις όλων των χρηστών για όλες τις ταινίες, επομένως πολλά διανύσματα αξιολόγησης θα είναι ελλιπή, με αποτέλεσμα η εκπαίδευση να μην μπορεί να λειτουργήσει, όταν απαντώνται τέτοιες τιμές. Μπορείτε να συμπληρώσετε τις ελλιπείς τιμές:

i) με μια τυχαία επιλεγμένη τιμή στο κατάλληλο εύρος,
ii) με μηδέν
iii) με τον μέσο όρο του διανύσματος αξιολόγησης.

Εξετάστε και τεκμηριώστε την επιλογή σας, σε σχέση και με το ζητούμενο (α).

γ) *Κανονικοποίηση (rescaling):* Με βάση την προτεινόμενη αρχιτεκτονική του ΤΝΔ, στην έξοδο αναμένονται τιμές αξιολογήσεων για έναν συγκεκριμένο χρήστη. Είναι προφανές ότι η συνάρτηση ενεργοποίησης στο επίπεδο εξόδου θα πρέπει να είναι σε θέση να παράξει τιμές σε αυτό το εύρος – ενδεχομένως με κάποιο γραμμικό μετασχηματισμό (βλ. και Α2-ε). Διαφορετικά, θα πρέπει τα δεδομένα να κανονικοποιηθούν σε διάστημα τιμών κατάλληλο για τη συγκεκριμένη συνάρτηση (π.χ. [0, 1] για τη λογιστική συνάρτηση).

δ) *Διασταυρούμενη Επικύρωση (cross-validation):* Βεβαιωθείτε ότι έχετε διαχωρίσει τα δεδομένα σας σε σύνολα εκπαίδευσης και ελέγχου, ώστε να χρησιμοποιήσετε 5-fold CV για όλα τα πειράματα. Παρατηρήστε ότι τέτοια σύνολα ήδη περιλαμβάνονται στο dataset (αρχεία u1.base έως u5.base και u1.test έως u5.test).

Α2. Επιλογή αρχιτεκτονικής [*45 μονάδες*]
Όσον αφορά την τοπολογία των ΤΝΔ για την εκπαίδευση τους με τον Αλγόριθμο Οπισθοδιάδοσης του Σφάλματος (back-propagation) θα χρησιμοποιήσετε ΤΝΔ με *ένα κρυφό επίπεδο* και θα πειραματιστείτε με τον αριθμό των κρυφών κόμβων. Για την εκπαίδευση του δικτύου χρησιμοποιήστε αρχικά ρυθμό μάθησης $\eta = 0.001$.

α) Η αξιολόγηση των μοντέλων σας μπορεί να γίνει με την *Ρίζα του Μέσου Τετραγωνικού Σφάλματος* (RMSE) καθώς και με το *Μέσο Απόλυτο Σφάλμα* (MAE) για τις αξιολογήσεις που

περιέχονται στα σύνολα ελέγχου[2]. Να εξηγήσετε με απλά λόγια ποια είναι η σημασία των παραπάνω μετρικών για το συγκεκριμένο πρόβλημα. [5]

β) Πόσες εισόδους θα χρειαστείτε στο ΤΝΔ, δεδομένου ότι μια είσοδος πρέπει να αναπαριστά έναν χρήστη; *Συμβουλή:* Για απλοποίηση, θεωρήστε ότι όλοι οι χρήστες αρχικά είναι ισοδύναμοι, επομένως ένας χρήστης μπορεί να αναπαρίσταται από ένα δυαδικό διάνυσμα Ν θέσεων, όσοι και οι χρήστες, όπου έχει 1 στη θέση που αντιστοιχεί στο χρήστη και 0 αλλού (τεχνική one-hot encoding). Να εξηγήστε γιατί είναι προτιμότερο να χρησιμοποιηθεί one-hot encoding για την κωδικοποίηση των χρηστών, έναντι π.χ. δυαδικής κωδικοποίησης. [5]

γ) Πόσους νευρώνες θα χρειαστείτε στο επίπεδο εξόδου, δεδομένου ότι η έξοδος πρέπει να αναπαριστά τις αξιολογήσεις του χρήστη για όλες τις ταινίες; [5]

δ) Να επιλέξετε κατάλληλη συνάρτηση ενεργοποίησης για τους κρυφούς κόμβους και να τεκμηριώσετε την επιλογή σας. [5]

ε) Ποια συνάρτηση ενεργοποίησης θα χρησιμοποιήσετε για το επίπεδο εξόδου; Σιγμοειδή, γραμμική ή κάποια άλλη; Εξετάστε την απάντησή σας σε σχέση με την κανονικοποίηση των αξιολογήσεων που κάνατε στο ζήτημα Α1. [5]

στ) Πειραματιστείτε με 3 διαφορετικές τιμές για τον αριθμό των νευρώνων του κρυφού επιπέδου και συμπληρώστε τον παρακάτω πίνακα, λαμβάνοντας υπόψη την υποσημείωση *1* (K αριθμός κόμβων στο κρυφό επίπεδο). Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ανά κύκλο εκπαίδευσης. Διατυπώστε τα συμπεράσματά σας σχετικά με τον αριθμό των κρυφών κόμβων. [15]

ζ) Κριτήριο τερματισμού. Επιλέξτε και τεκμηριώστε κατάλληλο κριτήριο τερματισμού της εκπαίδευσης κάθε φορά (για κάθε fold). Μπορεί να χρησιμοποιηθεί η τεχνική του *πρόωρου σταματήματος* (early stopping); [5]

Προσοχή: σε όλα τα πειράματα θα χρησιμοποιήσετε 5-fold cross validation (5-fold CV).

Αριθμός νευρώνων στο κρυφό επίπεδο	**RMSE**	**MAE**
H =		
H =		
H =		

Α3. Μεταβολές στο ρυθμό εκπαίδευσης και σταθεράς ορμής [20 *μονάδες*]
Επιλέγοντας την τοπολογία που δίνει το καλύτερο αποτέλεσμα βάσει του προηγούμενου ερωτήματος, να πραγματοποίησετε βελτιστοποίηση των υπερπαραμέτρων ρυθμού εκπαίδευσης η και σταθεράς ορμής m με χρήση CV και να συμπληρώστε τον παρακάτω πίνακα. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ως προς τους κύκλους εκπαίδευσης που θα χρειαστούν. Να τεκμηριώσετε θεωρητικά γιατί m < 1.

η	**m**	**RMSE**	**MAE**
0.001	0.2		
0.001	0.6		
0.05	0.6		
0.1	0.6		

[2] Παρατηρήστε ότι, στην περίπτωση ενός μόνο χρήστη, το RMSE ισοδυναμεί με την ευκλείδεια απόσταση του διανύσματος αξιολογήσεων που προβλέφθηκε από το πραγματικό, ενώ το MAE ισοδυναμεί με την απόσταση Manhattan.

Να διατυπώσετε σύντομα τα συμπεράσματα που προκύπτουν από τα 4 πειράματα.

Α4. Ομαλοποίηση [*15 μονάδες*]
Μια μέθοδος για την αποφυγή υπερπροσαρμογής του δικτύου και βελτίωση της γενικευτικής του ικανότητας είναι η ομαλοποίηση του διανύσματος των βαρών (regularization). Να εφαρμόσετε *L1* ομαλοποίηση και να επανεκπαιδεύσετε το δίκτυό σας, όπως προέκυψε από το Α3, αξιολογώντας διάφορες τιμές για τον συντελεστή φθοράς *r*. Γιατί είναι προτιμότερη η χρήση L1 ομαλοποίησης έναντι L2 στη συγκεκριμένη περίπτωση;

i) r = 0.1 ii) r = 0.5 iii) r = 0.9

Συμπληρώστε τον παρακάτω πίνακα για κάθε μία από τις παραπάνω περιπτώσεις με χρήση 5-fold CV. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ανά κύκλο εκπαίδευσης.

Συντελεστής φθοράς	**RMSE**	**MAE**
0.1		
0.5		
0.9		

Διατυπώστε τα συμπεράσματά σας σχετικά με την επίδραση της μεθόδου στην γενικευτική ικανότητα του δικτύου.

Α5. Βαθύ Νευρωνικό Δίκτυο. *[προαιρετικό ερώτημα - 10 μονάδες bonus]*
Δοκιμάστε να προσθέσετε περισσότερα του ενός κρυφά επίπεδα στο δίκτυο (μέχρι 3). Πειραματιστείτε με τον αριθμό των κόμβων, όπως κάνατε στο Α2. Περιγράψτε μια λογική για την στοίχιση των κρυφών επιπέδων (είναι καλό να έχουν τον ίδιο αριθμό κόμβων; Μειούμενο; Αυξανόμενο;). Να αναφέρετε RMSE και MAE για τα πειράματά σας με 5-fold CV.

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 επισυνάψετε και τον κώδικα, αν υπάρχει (όλα τα αρχεία θα τα συμπιέσετε σε ένα .zip αρχείο).
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, 20/4/2020, στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να επικοινωνείτε με τον κ. Ανδρέα Ανδριόπουλο, a.andriopoulos@upatras.gr