ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Διδάσκοντες: Σ. Λυκοθανάσης, Δ. Κουτσομητρόπουλος

Ακαδημαϊκό Έτος 2024-2025

Εργαστηριακή Άσκηση
Μέρος Α΄

Α. Πρόβλεψη Alzheimer's με Χρήση Νευρωνικών Δικτύων

Ένα ευρύ πεδίο εφαρμογής των αλγορίθμων και μοντέλων της υπολογιστικής νοημοσύνης είναι αυτό της βιοιατρικής. Με κατάλληλη εκπαίδευση, τα τεχνητά νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για την ταξινόμηση και πρόβλεψη της άνοιας, του Alzheimer και άλλων νευροεκφυλιστικών παθήσεων. Στην παρουσία εργασία σας ζητείται να σχεδιάσετε και να εκπαιδεύσετε ΤΝΔ με στόχο την πρόβλεψη Alzheimer από δεδομένα μετρήσεων βιοδεικτών, ιστορικού, αξιολογήσεων και εξετάσεων ανάμεσα σε δυνητικούς ασθενείς.

Για το σκοπό αυτό θα χρησιμοποιήσετε το σύνολο δεδομένων *Alzheimer's Disease Dataset*[1] που περιλαμβάνει δεδομένα για 2.149 ασθενείς. Το σύνολο δεδομένων διαθέτει 35 στήλες, εκ των οποίων η τελευταία αφορά τη διάγνωση Alzheimer και βρίσκεται στο αρχείο alzheimers_disease_data.csv[2]. Ένα απόσπασμα των δεδομένων αυτών φαίνεται στην παρακάτω εικόνα. Αναλυτική περιγραφή των χαρακτηριστικών και του εύρους τιμών δίνεται στο (1).

PatientID	Age	Gender	Ethnicity	Education...	BMI	Smoking	AlcoholCo...
4751	73	0	0	2	22.927749230993864	0	13.29721772827684
4752	89	0	0	0	26.827681191596 02	0	4.5425238177221905
4753	73	0	3	1	17.795882442817113	0	19.55508452555359
4754	74	1	0	1	33.800817044135 47	1	12.209265546203783
4755	89	0	0	0	20.716973826446807	0	18.454356090619612
4756	86	1	1	1	30.626885546270938	0	4.1401437842762 35
4757	68	0	3	2	38.387621858169126	1	0.6460472705489217

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, Keras, SciKit-Learn.*

Το ζητούμενο στην εργασία αυτή είναι να κατασκευαστεί και να εκπαιδευτεί ένα ΤΝΔ που να ταξινομεί εισόδους (χαρακτηριστικά ασθενών) σε μία από δύο κλάσεις (διάγνωση Alzheimer ή μη).

[1] R. E. Kharoua (2024). Alzheimer's Disease Dataset. https://www.kaggle.com/dsv/8668279
[2] https://eclass.upatras.gr/modules/document/file.php/CEID1060/alzheimers_disease_data.csv

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) *Κωδικοποίηση και προεπεξεργασία δεδομένων:* Στο σύνολο δεδομένων υπάρχουν ποσοτικές τιμές (π.χ. μετρήσεις βιοδεικτών), αλλά και κατηγορικές (π.χ. φύλο, εθνικότητα, κάπνισμα, ύπαρξη συμπτωμάτων κλπ) και διατεταγμένες τιμές (επίπεδο εκπαίδευσης, γνωστική κατάσταση, ποιότητα ύπνου κ.α.). Όλες οι τιμές είναι κωδικοποιημένες αριθμητικά, σε συνεχή, ακέραια ή δυαδική κλίμακα. Το εύρος τιμών των δεδομένων μπορεί να διαφέρει σημαντικά ανά χαρακτηριστικό. Για αυτό τον λόγο, υπάρχει κίνδυνος να υπερεκτιμηθεί η συνεισφορά κάποιου χαρακτηριστικού έναντι άλλων ή να υπάρξει πόλωση στις τιμές κάποιων χαρακτηριστικών ή στις τιμές των βαρών που θα προκύψουν. Με δεδομένη την συγκεκριμένη αποτύπωση και την πιθανή ανάγκη προσαρμογής των τιμών αυτών σε διαφορετική κλίμακα (εξάλειψη ενδεχόμενης πόλωσης), παρουσιάζονται οι εξής μέθοδοι:

- *Κεντράρισμα (Centering):* Με την μέθοδο αυτή αφαιρούμε τον μέσο όρο των τιμών για κάθε χαρακτηριστικό από όλες τις τιμές που έχουν αποδοθεί.
- *Κανονικοποίηση (Normalization ή min-max scaling):* Με την μέθοδο αυτή μεταφέρουμε το εύρος τιμών ενός χαρακτηριστικού σε νέα κλίμακα πχ [-1,1] ή [0, 1].
- *Τυποποίηση (Standardization ή z-score):* Με την μέθοδο αυτή παρέχουμε στο δείγμα ιδιότητες όπως μηδενική μέση τιμή και μοναδιαία διακύμανση (Gaussian).
- *One-hot encoding:* Κωδικοποιεί κατηγορικές τιμές σε δυαδικά διανύσματα με ακριβώς ένα 1 και 0 στις άλλες θέσεις.

Εξετάστε τη χρησιμότητα των ανωτέρω μεθόδων για το συγκεκριμένο πρόβλημα και εφαρμόστε τη/τις στα δεδομένα εκπαίδευσης, αν κρίνετε σκόπιμο. [15]

β) *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. Προσέξτε το κάθε fold να είναι *ισορροπημένο* (balanced) ως προς τον αριθμό των δειγμάτων κάθε κλάσης. [5]

Α2. Επιλογή αρχιτεκτονικής [*50 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 εκπαίδευση και αξιολόγηση των μοντέλων σας μπορεί να γίνει με χρήση *Cross-Entropy* (CE), *Μέσου Τετραγωνικό Σφάλματος* (MSE), καθώς και *ακρίβεια ταξινόμησης* (Accuracy)[3]. Να εξηγήσετε με απλά λόγια ποια είναι η σημασία των παραπάνω μετρικών για το συγκεκριμένο πρόβλημα. Ποια είναι προτιμότερη για εκπαίδευση (loss); [5]

β) Πόσους νευρώνες θα χρειαστείτε στο επίπεδο εξόδου, δεδομένου του ζητούμενου της ταξινόμησης σε δύο κλάσεις (binary classification); [5]

[3] Πρόκειται για το ποσοστό των ορθών προβλέψεων από το δίκτυο. Αν $\hat{y_i}$ είναι η επιθυμητή κλάση για το δείγμα i και y_i η κλάση που προβλέπει το δίκτυο, τότε $\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y_i} = y_i)$

γ) Να επιλέξετε κατάλληλη συνάρτηση ενεργοποίησης για τους κρυφούς κόμβους και να τεκμηριώσετε την επιλογή σας. Να αξιολογήσετε τουλάχιστον τη χρήση *Tanh, SiLU* και *ReLU*. [10]

δ) Ποια συνάρτηση ενεργοποίησης θα χρησιμοποιήσετε για το επίπεδο εξόδου; Σιγμοειδή, γραμμική, Softmax ή κάποια άλλη; [5]

ε) Πειραματιστείτε με 3 διαφορετικές τιμές για τον αριθμό των νευρώνων του κρυφού επιπέδου και συμπληρώστε τον παρακάτω πίνακα. Εμπειρικά ενδεδειγμένες τιμές για τον αριθμό των κρυφών κόμβων *H* βρίσκονται στο διάστημα [*I/2, 2I*] (*I αριθμός εισόδων, H αριθμός κόμβων στο κρυφό επίπεδο)*. Να συμπεριλάβετε και τις γραφικές παραστάσεις σύγκλισης (Μ.Ο.) ανά κύκλο εκπαίδευσης.
Διατυπώστε τα συμπεράσματά σας σχετικά με (i) τον αριθμό των κρυφών κόμβων, (ii) την επιλογή της συνάρτησης κόστους, (iii) την επιλογή της συνάρτησης ενεργοποίησης των κρυφών κόμβων και (iv) την ταχύτητα σύγκλισης ως προς τις εποχές εκπαίδευσης. [20]

Αριθμός νευρώνων στο κρυφό επίπεδο	**CE loss**	**MSE**	**Acc**
$H_1 = I/2$			
$H_1 = 2I/3$			
$H_1 = I$			
$H_1 = 2I$			

στ) 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).

Α3. Μεταβολές στον ρυθμό εκπαίδευσης και σταθεράς ορμής [15 *μονάδες*]
Επιλέγοντας την τοπολογία που δίνει το καλύτερο αποτέλεσμα βάσει του προηγούμενου ερωτήματος, να πραγματοποιήσετε 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. Να διατυπώσετε σύντομα τα συμπεράσματα που προκύπτουν από τα 4 πειράματα.

η	**m**	**CE loss**	**MSE**	**Acc**
0.001	0.2			
0.001	0.6			
0.05	0.6			
0.1	0.6			

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 εξηγήσετε ποια μέθοδο ομαλοποίησης (*L1* ή *L2*) είναι προτιμότερη για το συγκεκριμένο πρόβλημα. Στη συνέχεια να την εφαρμόσετε και να επανεκπαιδεύσετε το δίκτυό σας, όπως προέκυψε από το Α3, αξιολογώντας διάφορες τιμές για τον συντελεστή *r*.

i) r = 0.0001 ii) r = 0.001 iii) r = 0.01

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

Συντελεστής *r*	**CE loss**	**MSE**	**Acc**
0.0001			
0.001			
0.01			

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 CE, MSE και Acc για τα πειράματά σας με 5-fold CV και να διατυπώσετε τα συμπεράσματά σας σχετικά με την προσθήκη κρυφών επιπέδων.

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 link προς τον κώδικα που έχετε χρησιμοποιήσει (σε κάποια file sharing υπηρεσία ή code repo).
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 Μ. Δευτέρα, 14/4/2025, στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να χρησιμοποιείτε το σχετικό forum στο eclass του μαθήματος.