Εργασία 4 - Υπολογιστική Νοημοσύνη
Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK

Στόχος της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των μοντέλων TSK στην επίλυση προβλημάτων ταξινόμησης (classification). Συγκεκριμένα, επιλέγονται δύο σύνολα δεδομένων από το UCI repository με σκοπό την ταξινόμηση, από τα διαθέσιμα δεδομένα, δειγμάτων στις εκάστοτε κλάσεις τους, με χρήση ασαφών νευρωνικών μοντέλων. Η εργασία αποτελείται από δύο μέρη, το πρώτο από τα οποία προορίζεται για μια απλή διερεύνηση της διαδικασίας εκπαίδευσης και αξιολόγησης των TSK μοντέλων, ενώ το δεύτερο περιλαμβάνει μια πιο συστηματική προσέγγιση στο πρόβλημα της εκμάθησης από δεδομένα, σε συνδυασμό με προεπεξεργαστικά βήματα όπως επιλογή χαρακτηριστικών (feature selection) και μεθόδους βελτιστοποίησης των μοντέλων μέσω της διασταυρωμένης επικύρωσης (cross validation).

1 Εφαρμογή σε απλό dataset

Στην πρώτη φάση της εργασίας, επιλέγεται από το UCI repository το Haberman's Survival, το οποίο περιλαμβάνει 306 δείγματα (instances), από 3 χαρακτηριστικά (attributes) το καθένα. Ακολουθούνται τα εξής βήματα:

- *Διαχωρισμός σε σύνολα εκπαίδευσης-επικύρωσης-ελέγχου*: Σε πρώτη φάση είναι απαραίτητος ο διαχωρισμός του συνόλου δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα $D_{trn}, D_{val}, D_{chk}$, από τα οποία το πρώτο θα χρησιμοποιηθεί για εκπαίδευση, το δεύτερο για επικύρωση και αποφυγή του φαινομένου υπερεκπαίδευσης και το τελευταίο για τον έλεγχο της απόδοσης του τελικού μας μοντέλου. Προτείνεται να χρησιμοποιηθεί το 60% του συνόλου των δειγμάτων για το υποσύνολο εκπαίδευσης και από 20% του συνόλου των δειγμάτων για κάθε ένα από τα δύο εναπομείναντα υποσύνολα. Ένα σημείο στο οποίο θα πρέπει να δοθεί προσοχή είναι το ότι για να επιτύχουμε καλή απόδοση, θα πρέπει η συχνότητα εμφάνισης δειγμάτων που ανήκουν σε μια συγκεκριμένη κλάση, σε κάθε ένα από τα τρία σύνολα διαμέρισης, να είναι όσο το δυνατόν πιο "όμοια" με την αντίστοιχη συχνότητα εμφάνισής τους στο αρχικό σύνολο δεδομένων.

- *Εκπαίδευση TSK μοντέλων με διαφορετικές παραμέτρους:* Σε αυτό το στάδιο θα εξεταστούν διάφορα μοντέλα TSK όσον αφορά την απόδοσή τους στο σύνολο ελέγχου. Συγκεκριμένα, θα εκπαιδευτούν τέσσερα TSK μοντέλα, στα οποία θα μεταβάλλεται το πλήθος των ασαφών IF-THEN κανόνων.

	Actual: C_1	Actual: C_2	...	Actual: C_k
Predicted: C_1	x_{11}	x_{12}	...	x_{1k}
Predicted: C_2	x_{21}	x_{22}	...	x_{2k}
...	...	...	...	...
Predicted: C_k	x_{k1}	x_{k2}	...	x_{kk}

Πίνακας 1: Error matrix

Σκοπός είναι να μελετηθεί η επίδραση της διαμέρισης του χώρου εισόδου – σε συνάρτηση με την πολυπλοκότητα που αυτή επιφέρει, στην απόδοση του ταξινομητή. Η διαμέριση του χώρου εισόδου θα γίνει με τη μέθοδο του Subtractive Clustering (SC) και τα TSK μοντέλα που θα προκύψουν θα διαφέρουν ως προς την παράμετρο που καθορίζει τον αριθμό των κανόνων. Στην πρώτη περίπτωση (δύο πρώτα μοντέλα), το subtractive clustering θα εκτελεστεί για όλα τα δεδομένα του συνόλου εκπαίδευσης (class independent), ενώ στη δεύτερη (επόμενα δύο), θα εξεταστεί ο διαμερισμός του χώρου εισόδου εφαρμόζοντας clustering στα δεδομένα του συνόλου εκπαίδευσης που ανήκουν στην εκάστοτε κλάση ξεχωριστά (class dependent). Ο λόγος για τον οποίο γίνεται αυτό είναι η αύξηση της ερμηνευσιμότητας του μοντέλου και η παραγωγή ¨καθαρότερων' clusters (άρα και κανόνων).

Σημείωση: Εφόσον η έξοδός μας αποτελείται από έναν ακέραιο αριθμό, ενδεικτικό της κλάσης στην οποία ανήκει το εκάστοτε δείγμα, προτείνεται η χειροκίνητη αλλαγή του τύπου συνάρτησης εξόδου από γραμμική σε singleton. Για τις δύο περιπτώσεις που θα εξεταστούν (class independent - class dependent) η παράμετρος που καθορίζει το μέγεθος των clusters (και τον αριθμό των κανόνων) να λάβει δύο ακραίες τιμές, έτσι ώστε ο αριθμός των κανόνων στα δύο μοντέλα που θα προκύψουν να παρουσιάζει σημαντική διαφορά. Και τα τέσσερα μοντέλα να εκπαιδευτούν με την υβριδική μέθοδο, σύμφωνα με την οποία οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται μέσω της μεθόδου της οπισθοδιάδοσης (backpropagation algorithm), ενώ οι παράμετροι της συνάρτησης εξόδου βελτιστοποιούνται μέσω της μεθόδου των ελαχίστων τετραγώνων (Least Squares). Αξιολόγηση μοντέλων: Για την αξιολόγηση της ταξινόμησης των δειγμάτων από τα διάφορα μοντέλα,

- *Αξιολόγηση μοντέλων:* Για την αξιολόγηση της ταξινόμησης των δειγμάτων από τα διάφορα μοντέλα, θα χρησιμοποιηθούν οι εξής δείκτες απόδοσης:

  1. *Error matrix:* Ο πίνακας σφαλμάτων ταξινόμησης είναι ένας $k \times k$ πίνακας, με k τον αριθμό των κλάσεων ο οποίος βοηθά στην οπτικοποίηση της απόδοσης ενός ταξινομητή και μέσω του οποίου αποκτούμε πρόσβαση σε μια σειρά δεικτών απόδοσης. Η γενική του δομή παρουσιάζεται στον πίνακα 1.

     Τα στοιχεία της κύριας διαγωνίου περιλαμβάνουν το πλήθος των δειγμάτων που ανήκουν σε μια συγκεκριμένη κλάση και τα οποία ορθώς

ταξινομήθηκαν σε αυτή από το μοντέλο μας, ενώ τα στοιχεία εκτός της διαγωνίου περιλαμβάνουν το πλήθος των δειγμάτων τα οποία λανθασμένα ταξινομήθηκαν σε διαφορετική κλάση από αυτή στην οποία στην πραγματικότητα ανήκουν.

2. *Overall accuracy:* Η συνολική ακρίβεια ενός ταξινομητή ορίζεται ως το ποσοστό των ορθώς ταξινομημένων δειγμάτων ως προς το συνολικό πλήθος των δειγμάτων. Χρησιμοποιώντας τα στοιχεία του πίνακα σφαλμάτων, η ακρίβεια υπολογίζεται ως:

$$OA = \frac{1}{N}\sum_{i=1}^{k} x_{ii}$$

3. *3. Producer's accuracy – User's accuracy:* ύο δείκτες που παρουσιάζουν ενδιαφέρον και που αναφέρονται στην απόδοση του ταξινομητή όσον αφορά κάθε κλάση ξεχωριστά, είναι η ακρίβεια παραγωγού και η ακρίβεια χρήστη. Ορίζουμε αρχικά $x_{ir} = \sum_{j=1}^{k} x_{ij}$ ο πλήθος των σημείων που ταξινομήθηκαν στην κλάση C_i και $x_{jc} = \sum_{i=1}^{k} x_{ij}$ το πλήθος των σημείων τα οποία ανήκουν στην κλάση C_j. Με βάση τα παραπάνω, η ακρίβεια παραγωγού δίνεται από τον τύπο $PA(j) = \frac{x_{jj}}{x_{jc}}$ και η ακρίβεια χρήστη θα είναι $UA(i) = \frac{x_{ii}}{x_{ir}}$

4. $\hat{K}$: Ένα άλλο στατιστικό μέγεθος που μπορεί να εξαχθεί από έναν πίνακα σφαλμάτων είναι το μέγεθος $\hat{\kappa}$, το οποίο αποτελεί εκτίμηση της πραγματικής στατιστικής παραμέτρου . Υπολογίζεται σύμφωνα με τον τύπο

$$\hat{\kappa} = \frac{N\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{ir}x_{ic}}{N^2 - \sum_{i=1}^{k} x_{ic}x_{ir}}$$

- *Ζητούμενα του προβλήματος:* Για κάθε ένα από τα πέντε TSK μοντέλα, να γίνουν οι κατάλληλες αρχικοποιήσεις και στη συνέχεια να εκτελεστεί η εκπαίδευσή τους με τις παραμέτρους που περιγράφηκαν παραπάνω. Ζητούνται τα εξής:

  1. Να δώσετε τα αντίστοιχα διαγράμματα στα οποία να απεικονίζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν μεσω της διαδικασίας εκπαίδευσης.
  2. Να δοθούν τα διαγράμματα μάθησης (learning curves) όπου να απεικονίζεται το σφάλμα του μοντέλου συναρτήσει του αριθμού των επαναλήψεων (iterations).
  3. Να δοθεί ο πίνακας σφαλμάτων ταξινόμησης και να εξαχθούν από αυτόν τιμές των δεικτών απόδοσης $OA, PA, UA, \hat{\kappa}$
  4. Να σχολιάσετε τα αποτελέσματα. Ποιά είναι η επίδραση του αριθμού των κανόνων στην απόδοση του ταξινομητή; Ποιά συμπεράσματα μπορούμε να εξάγουμε σχετικά με την επικάλυψη των προβολών των ασαφών συνόλων κάθε cluster στις αντίστοιχες εισόδους όσον αφορά

την ενεργοποίηση των κανόνων και γενικότερα την απόδοση του ταξινομητή· Να συνοδεύσετε τα σχόλια με διαγράμματα της επιλογής σας. Μπορείτε να προτείνετε κάποια μέθοδο για τη βελτίωση της σχεδίασης του τμήματος υπόθεσης·

Σημείωση για την έξοδο των μοντέλων: Ένα σημείο το οποίο μπορεί να αποτελέσει πηγή σύγχυσης, είναι το γεγονός ότι η υλοποίηση των TSK ασαφών μοντέλων στο MATLAB είναι τέτοια ώστε η έξοδός τους να είναι πραγματική, κάτι το οποίο οδηγεί σε δυσκολίες σε προβλήματα ταξινόμησης, όπου η μεταβλητή – στόχος είναι συνήθως κατηγορική. Στα προβλήματα ταξινόμησης που συναντούμε σε αυτή την εργασία, η μεταβλητή – στόχος είναι ακέραιος,λαμβάνοντας τιμές σε κάποιο σύνολο ακεραίων $\{k_0, k_1, \ldots, k_k\}$. Ένας απλός τρόπος να φέρουμε την έξοδο του μοντέλου στην ίδια μορφή είναι να στρογγυλοποιήσουμε κάθε στοιχείο στον πλησιέστερο ακέραιο. Εναλλακτικά, μπορείτε να ορίσετε ένα δικό σας σχήμα διακριτοποίησης της συνεχούς εξόδου του μοντέλου, αν κρίνετε ότι κάτι τέτοιο οδηγεί σε αποδοτικότερη ταξινόμηση.

2 Εφαρμογή σε dataset με υψηλή διαστασιμότητα:

Στη δεύτερη φάση της εργασίας θα ακολουθηθεί μια πιο συστηματική προσέγγιση στο πρόβλημα της χρήσης ασαφών νευρωνικών μοντέλων σε προβλήματα ταξινόμησης. Για το σκοπό αυτό θα επιλεχθεί ένα dataset με υψηλότερο βαθμό διαστασιμότητας. Ένα προφανές πρόβλημα που ανακύπτει από την επιλογή αυτή, είναι η λεγόμενη "έκρηξη" του πλήθους των IF-THEN κανόνων (rule explosion). Όπως είναι γνωστό από τη θεωρία, για την κλασική περίπτωση του γριδ παρτιτιονινγ του χώρου εισόδου, ο αριθμός των κανόνων αυξάνεται εκθετικά σε σχέση με το πλήθος των εισόδων, γεγονός που καθιστά πολύ δύσκολη την μοντελοποίηση μέσω ενός TSK μοντέλου ακόμα και για datasets μεσαίας κλίμακας.

Το dataset που θα επιλεγεί για την επίδειξη των παραπάνω μεθόδων είναι το Epileptic Seizure Recognition dataset από το UCI repository. Το συγκεκριμένο dataset, περιλαμβάνει 11500 δείγματα, καθένα από τα οποία περιγράφεται από 179 μεταβλητές/χαρακτηριστικά. Είναι φανερό ότι το μέγεθος του dataset καθιστά δυσκολότερη μια απλή εφαρμογή ενός TSK μοντέλου, σαν αυτή του προηγούμενου μέρους της εργασίας. Ο μεγάλος αριθμός μεταβλητών καθιστά αναγκαία τη χρήση μεθόδων μείωσης της διαστασιμότητας καθώς και του αριθμού των IF-THEN κανόνων. Ο στόχος αυτός θα επιτευχθεί μέσω της επιλογής χαρακτηριστικών και της χρήσης ασαφούς ομαδοποίησης. Οι δύο αυτές μέθοδοι όμως, παρά τη ελάττωση της πολυπλοκότητας που επιφέρουν, εισάγουν στο πρόβλημα δύο ελεύθερες παραμέτρους, συγκεκριμένα, τον αριθμό των χαρακτηριστικών προς επιλογή και τον αριθμό των ομάδων που θα δημιουργηθούν. Η επιλογή των δύο αυτών παραμέτρων επαφίεται στον εκάστοτε χρήστη και είναι ουσιαστική όσον αφορά την τελική απόδοση του μοντέλου. Στην παρούσα εργασία, θα υλοποιηθεί η μέθοδος αναζήτησης πλέγματος για την εύρεση των βέλτιστων τιμών των παραμέτρων. Αναλυτικά, η μοντελοποίηση του προβλήματος θα ακολουθήσει τα εξής βήματα:

1. *Διαχωρισμός σε σύνολα εκπαίδευσης- επικύρωσης – ελέγχου:* Όπως και στο πρώτο κομμάτι της εργασίας, είναι απαραίτητος ο διαχωρισμός του συνόλου δεδομένων σε τρία υποσύνολα $D_{trn}, D_{val}, D_{chk}$, το ένα από τα οποία θα χρησιμοποιηθεί για εκπαίδευση και το δεύτερο για έλεγχο της απόδοσης.

2. *Επιλογή των βέλτιστων παραμέτρων:* Όπως αναφέρθηκε παραπάνω, το σύστημά μας περιλαμβάνει δύο ελεύθερες παραμέτρους την τιμή των οποίων πρέπει να επιλέξουμε εμείς. Η δημοφιλέστερη μέθοδος μέσω της οποίας επιτυγχάνεται αυτό είναι η αναζήτηση πλέγματος. Συγκεκριμένα, αφού λάβουμε ένα σύνολο τιμών για κάθε παράμετρο, δημιουργούμε ένα n-διάστατο πλέγμα (στην περίπτωσή μας $n = 2$), όπου κάθε σημείο αντιστοιχεί σε μια n-άδα τιμών για τις εν λόγω παραμέτρους, και σε κάθε σημείο χρησιμοποιούμε μια μέθοδο αξιολόγησης για ελέγξουμε την ορθότητα των συγκεκριμένων τιμών. Μια καθιερωμένη επιλογή για την αξιολόγηση αυτή αποτελεί η διασταυρωμένη επικύρωση (cross validation). Σύμφωνα με τη μέθοδο αυτή ,και για επιλεγμένες τιμές των παραμέτρων, χωρίζουμε το σύνολο εκπαίδευσης σε δύο υποσύνολα, από τα οποία το ένα θα χρησιμοποιηθεί για την εκπαίδευση ενός μοντέλου και το δεύτερο για την αξιολόγησή του. Η διαδικασία αυτή επαναλαμβάνεται – συνήθως πέντε ή δέκα φορές – όπου κάθε φορά χρησιμοποιείται διαφορετικός διαχωριμός του συνόλου εκπαίδευσης, και στο τέλος λαμβάνουμε τον μέσο όρο του σφάλματος του μοντέλου. Η λογική πίσω από τις πολλαπλές εκπαιδεύσεις και ελέγχους έγκειται στο ότι με αυτό τον τρόπο, αποκτούμε μια αρκετά καλή εκτίμηση της απόδοσης του μοντέλου, και έμμεσα των τιμών των παραμέτρων με βάση τις οποίες χτίστηκε το μοντέλο. Όταν η παραπάνω διαδικασία εκτελεστεί για κάθε σημείο του πλέγματος, λαμβάνουμε ως βέλτιστες τιμές των παραμέτρων, τις τιμές που αντιστοιχούν στο μοντέλο που παρουσίασε το ελάχιστο μέσο σφάλμα. Οι τιμές αυτές χρησιμοποιούνται για την εκπαίδευση του τελικού μας μοντέλου.

   Για τους σκοπούς της εργασίας, ορίζουμε τις εξής παραμέτρους:

   - Αριθμός χαρακτηριστικών: Το πλήθος των χαρακτηριστικών που θα χρησιμοποιηθούν στην εκπαίδευση των μοντέλων.
   - Ακτίνα των clusters r_α: Η παράμετρος που καθορίζει την ακτίνα επιρροής των clusters και κατ΄ επέκταση το πλήθος των κανόνων που θα προκύψουν.

   Ο καθορισμός των τιμών των παραμέτρων που θα εξεταστούν επιλέγεται ελεύθερα.

3. Με βάση τις βέλτιστες τιμές των παραμέτρων που επιλέχθηκαν από το προηγούμενο βήμα, εκπαιδεύουμε ένα τελικό TSK μοντέλο και ελέγχουμε την απόδοσή του στο σύνολο ελέγχου.

Τα παραπάνω βήματα συνοψίζουν πλήρως τη διαδικασία μοντελοποίησης που θα ακολουθηθεί. Ζητούνται τα εξής:

1. Ο διαχωρισμός του συνόλου δεδομένων να γίνει όπως και στο πρώτο κομμάτι, με τα σύνολα εκπαίδευσης-επικύρωσης-ελέγχου να περιλαμβάνουν αντίστοιχα το 60% - 20% - 20% του συνόλου.

2. Να εκτελεστεί αναζήτηση πλέγματος (grid search) και αξιολόγηση μέσω 5-πτυχης διασταυρωμένης επικύρωσης (5-fold cross validation) για την επιλογή των βέλτιστων τιμών των παραμέτρων. Σε κάθε επανάληψη να αποθηκεύεται το μέσο σφάλμα. Ο διαχωρισμός των δεδομένων να γίνει έτσι ώστε σε κάθε επανάληψη, το 80% των δεδομένων να χρησιμοποιείται για εκπαίδευση και το υπόλοιπο 20% για επικύρωση (ως είσοδοι στη συνάρτηση anfis του MATLAB). Μια συνάρτηση που μπορεί να βοηθήσει σε αυτό το έργο είναι η cvpartition. Θα πρέπει κι εδώ να δοθεί προσοχή έτσι ώστε η κατανομή των κλάσεων να διατηρηθεί και στα δύο υποσύνολα. Ως αλγόριθμος επιλογής χαρακτηριστικών μπορεί να επιλεχθεί ένας από τους εξής: (Relief, mRMR, FMI) και ως μέθοδος διαμέρισης διασκορπισμού ο αλγόριθμος Subtractive Clustering (SC). Να εφαρμοστεί προεπεξεργασία των δεδομένων αν αυτό κριθεί απαραίτητο. Μετά το πέρας της διαδικασίας, να σχολιαστούν τα αποτελέσματα όσον αφορά το μέσο σφάλμα σε συνάρτηση με τις τιμές των παραμέτρων. Να δοθούν διαγράμματα τα οποία να απεικονίζουν την καμπύλη αυτού του σφάλματος σε σχέση με τον αριθμό των κανόνων και σε σχέση με τον αριθμό των επιλεχθέντων χαρακτηριστικών. Ποιά συμπεράσματα μπορούν να βγουν·

3. Να εκπαιδευτεί το τελικό TSK μοντέλο με τις βέλτιστες τιμές των παραμέτρων. Να δοθούν τα εξής διαγράμματα:

   - Διαγράμματα όπου να αποτυπώνονται οι προβλέψεις του τελικού μοντέλου καθώς και οι πραγματικές τιμές.
   - Διαγράμματα εκμάθησης όπου να απεικονίζεται το σφάλμα συναρτήσει του αριθμού επαναλήψεων.
   - Να δοθούν ενδεικτικά μερικά ασαφή σύνολα στην αρχική και τελική τους μορφή.
   - Να δοθεί ο πίνακας σφαλμάτων ταξινόμησης και να εξαχθούν από αυτόν τιμές των δεικτών απόδοσης $OA, PA, UA, \hat{\kappa}$
   - Τέλος, να σχολιαστούν τα αποτελέσματα όσον αφορά τα χαρακτηριστικά που επιλέχθηκαν και τον αριθμό IF-THEN κανόνων του ασαφούς συστήματος συμπερασμού. Να γίνει σύγκριση με τον αντίστοιχο αριθμό κανόνων αν για το ίδιο πλήθος χαρακτηριστικών, είχαμε επιλέξει grid partitioning με δύο ή τρία ασαφή σύνολα ανά είσοδο. Ποιά είναι τα συμπεράσματα· Τέλος, να γίνουν αντίστοιχα σχόλια όπως και στο πρώτο τμήμα, σχετικά με την επικάλυχη των προβολών των ασαφών συνόλων στο χώρο των μεταβλητών εισόδου και την επίδραση του διαμερισμού του συνολικού χώρου εισόδου στο ποσοστό των ενεργών κανόνων.

Σημείωση:

1. Όπως και στο πρώτο κομμάτι της εργασίας, να γίνει χειροκίνητη αλλαγή του τύπου της συνάρτησης εξόδου των μοντέλων από linear σε constant.

2. Η διαδικασία του subtractive clustering θα πρέπει να εφαρμοστεί ξεχωριστά για κάθε κλάση (με την ίδια τιμή στην παράμετρο που καθορίζει την ακτίνα.). Μετά τον αρχικό διαχωρισμό των δεδομένων, τα δεδομένα εκπαίδευσης διαμερίζονται επιπλέον ανάλογα με την κλάση στην οποία ανήκουν και η συνολική βάση κανόνων προκύπτει ως η ένωση των επιμέρους κανόνων που προκύπτουν από τις class-specific ομαδοποιήσεις. Το σχετικό παράδειγμα που έχει ανέβει στο e-learning παρέχει μια επίδειξη αυτής της διαδικασίας.