Εργασία 3 - Υπολογιστική Νοημοσύνη
Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Στόχος της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των μοντέλων TSK στη μοντελοποίηση πολυμεταβλητών, μη γραμμικών συναρτήσεων. Συγκεκριμένα, επιλέγονται δύο σύνολα δεδομένων από το UCI repository με σκοπό την εκτίμηση της μεταβλητής στόχου από τα διαθέσιμα δεδομένα, με χρήση ασαφών νευρωνικών μοντέλων. Το πρώτο σύνολο δεδομένων θα χρησιμοποιηθεί για μια απλή διερεύνηση της διαδικασίας εκπαίδευσης και αξιολόγησης μοντέλων αυτού του είδους, καθώς και για μια επίδεξη τρόπων ανάλυσης και ερμηνείας των αποτελεσμάτων. Το δεύτερο, πολυπλοκότερο σύνολο δεδομένων θα χρησιμοποιηθεί για μια πληρέστερη διαδικασία μοντελοποίησης, η οποία θα περιλαμβάνει μεταξύ άλλων προεπεξεργαστικά βήματα όπως επιλογή χαρακτηριστικών (feature selection), καθώς 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 Airfoil Self-Noise dataset, το οποίο περιλαμβάνει 1503 δείγματα (instances) και 6 χαρακτηριστικά (features). Ακολουθούμε 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.

- *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 4 TSK μοντέλα, στα οποία θα μεταβάλλονται η μορφή της εξόδου καθώς και το πλήθος των συναρτήσεων συμμετοχής για κάθε μεταβλητή εισόδου. Δίνεται ο Πίνακας 1 για περαιτέρω επεξήγηση. Και τα 4 μοντέλα θα εκπαιδευτούν με την υβριδική μέθοδο, σύμφωνα με την οποία οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται μέσω της μεθόδου της οπισθοδιάδοσης (backpropagation

	Πλήθος συναρτήσεων συμμετοχής	Μορφή εξόδου
TSK_model_1	2	Singleton
TSK_model_2	3	Singleton
TSK_model_3	2	Polynomial
TSK_model_4	3	Polynomial

Πίνακας 1: Ταξινόμηση μοντέλων προς εκπαίδευση.

algorithm), ενώ οι παράμετροι της πολυωνυμικής συνάρτησης εξόδου βελτιστοποιούνται μέσω της μεθόδου των ελαχίστων τετραγώνων (Least Squares). Οι συναρτήσεις συμμετοχής να είναι bell-shaped και η αρχικοποίησή τους να γίνει με τέτοιον τρόπο ώστε τα διαδοχικά ασαφή σύνολα να παρουσιάζουν σε κάθε είσοδο, βαθμό επικάλυψης περίπου 0.5.

- *Αξιολόγηση μοντέλων:* Για την ακρίβεια της εκτίμησης της πραγματικής συνάρτησης από κάθε ένα από τα παραπάνω μοντέλα, θα χρησιμοποιηθούν οι εξής δείκτες απόδοσης:

  1. *MSE*: ο μέσο τετραγωνικό σφάλμα μεταξύ της εξόδου του μοντέλου και της πραγματικής εξόδου, όπου το διάνυσμα παραμέτρων του τελικού μοντέλου που έχει επιλεχθεί.

  $$MSE(\boldsymbol{\theta}) = \sigma_e^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N}(f(\boldsymbol{x}_i) - \hat{f}(\boldsymbol{x}_i))^2$$

  Εναλλακτικά, μπορεί να χρησιμοποιηθεί ο δείκτης RMSE, ο οποίος είνα απλά η τετραγωνική ρίζα του δείκτη MSE:

  $$RMSE(\boldsymbol{\theta}) = \sqrt{MSE(\boldsymbol{\theta})}$$

  2. *Συντελεστής προσδιορισμού R^2:* 2. Ο συντελεστής προσδιορισμού R^2, ο οποίος μας δίνει πληροφορία για το ποσοστό της διακύμανσης της πραγματικής το οποίο "εξηγείται" από το μοντέλο μας. Ο υπολογισμός του συντελεστή προσδιορισμού μπορεί να γίνει με τον εξής τρόπο:

  $$SS_{res} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$
  $$SS_{tot} = \sum_{i=1}^{N}(y_i - \bar{y})^2$$
  $$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

  όπου y η είναι η πραγματική έξοδος του συστήματος, $\hat{y}$ η εκτίμηση που παράγει το μοντέλο μας και $\bar{y}$ η μέση τιμή της πραγματικής εξόδου.

3. Τέλος, για πιο εύρωστη αξιολόγηση, θα μπορούσαν να χρησιμοποιηθούν οι δείκτες NMSE και NDEI οι οποίοι υπολογίζονται ως εξής:

$$NMSE = \frac{\sigma_e^2}{\sigma_x^2} = \frac{\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$
$$NDEI = \sqrt{NMSE} = \frac{\sigma_e}{\sigma_x}$$

- *Ζητούμενα του προβλήματος*: Για κάθε ένα από τα 4 TSK μοντέλα που περιγράφονται στον παραπάνω πίνακα, να γίνουν οι κατάλληλες αρχικοποιήσεις και στη συνέχεια να εκτελεστεί η εκπαίδευση των μοντέλων με τις παραμέτρους που περιγράφηκαν παραπάνω. Ως τελικό μοντέλο να επιλέγεται πάντα εκείνο το οποίο αντιστοιχεί στο μικρότερο σφάλμα στο σύνολο επικύρωσης. Για τις τέσσερις περιπτώσεις εκπαίδευσης:

  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 δοθούν τα διαγράμματα όπου να αποτυπώνονται τα σφάλματα πρόβλεψης.
  4. Τέλος, να παρουσιαστούν σε μορφή πίνακα οι τιμές των δεικτών απόδοσης $RMSE, NMSE, NDEI, R^2$

Να σχολιάσετε τα αποτελέσματα των μοντέλων τόσο όσον αφορά τη μορφή της εξόδου όσο και την διαμέριση του χώρου εισόδου. Το μεγαλύτερο πλήθος ασαφών συνόλων ανά είσοδο στην περίπτωση των αντίστοιχων TSK μοντέλων οδήγησε σε υπερεκπαίδευση· Να ερμηνευτούν οποιεσδήποτε διαφορές στην απόδοση των τεσσάρων μοντέλων.

2 Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Στη δεύτερη φάση της εργασίας θα ακολουθηθεί μια πιο συστηματική προσέγγιση στο πρόβλημα μοντελοποίησης μιας άγνωστης συνάρτησης. 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 Superconductivty dataset από το UCI Repository, το οποίο περιλαμβάνει 21263 δείγματα καθένα από τα οποία περιγράφεται από 81 μεταβλητές/χαρακτηριστικά. Είναι φανερό ότι το μέγεθος του δαταset καθιστά απαγορευτική μια απλή εφαρμογή ενός ΤΣΚ 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 (π.χ. με 81 μεταβλητές/predictors, διαμερίζαμε το χώρο εισόδου κάθε μεταβλητής με δύο ασαφή σύνολα, θα καταλήγαμε με 2^{81} κανόνες). Ο στόχος αυτός θα επιτευχθεί μέσω της επιλογής χαρακτηριστικών και της χρήσης διαμέρισης διασκορπισμού. 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 λοιπόν 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. Σημειώνεται ότι τα σύνολα παραμέτρων έτσι όπως παρουσιάζονται παραπάνω είναι προαιρετικά, και μπορεί κανείς να αντικαταστήσει τις τιμές τους, ειδικά αν η διαδικασία της διασταυρωμένης επικύρωσης αποδειχθεί ιδιαίτερα χρονοβόρα. 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. 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). Ως μέθοδος ομαδοποίησης για τη δημιουργία των IF-THEN κανόνων επιλέγεται ο αλγόριθμος Subtractive Clustering (SC) και η επιλογή χαρακτηριστικών μπορεί να εκτελεστεί με έναν από τους εξής αλγορίθμους (Relief, mRMR, FMI). 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 και με τις ίδιες προδιαγραφές όπως και προηγουμένως (SC). 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 δοθούν σε ένα πίνακα οι τιμές των δεικτών απόδοσης $RMSE, NMSE, NDEI, R^2$
- 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·