Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Μελέτη και υλοποίηση

Εργασία #3 στο μάθημα

της Υπολογιστικής Νοημοσύνης

του

ΒΑΣΙΛΕΙΟΥ ΠΟΛΥΝΟΠΟΥΛΟΥ (ΑΕΜ:9584)

Διδάσκοντες: Θεοχάρης Ιωάννης
Χαδουλός Χρήστος

Θεσσαλονίκη, Μάιος 2024

Περιεχόμενα

Γενική Περιγραφή

Η εργασία αυτή έχει σκοπό τη διερεύνηση της ικανότητας μοντέλων TSK στη μοντελοποίηση πολυμεταβλητών, μη γραμμικών συναρτήσεων και απαρτίζεται από δύο εφαρμογές που λαμβάνουν δεδομένα από μικρής και υψηλής διαστασιμότητας dataset αντίστοιχα. Η πρώτη εφαρμογή δεχεται το μικρό σύνολο δεδομένων για τη ανάλυση της διαδικασίας εκπαίδευσης, αξιολόγησης των μοντέλων και ερμηνείας των αποτελεσμάτων. Η δεύτερη εφαρμογή χρησιμοποιεί μια πληρέστερη διαδικασία μοντελοποίησης, υλοποιώντας προεπεξεργαστικά βήματα όπως η επιλογή χαρακτηριστικών (feature selection) και λαμβάνοντας μεθόδους βελτιστοποίησης όπως η διασταυρωμένη επικύρωση (cross validation).

Μέρος Α : Εφαρμογή σε απλό dataset

Δεδομένα και Ζητούμενα

Το πρωτο μέρος της εργασίας λαμβάνει το απλό *Airfoil Self-Noise dataset* από το *UCI repository* με 1503 δείγματα (instances) και 6 χαρακτηριστικά (features). Γίνεται διαχωρισμός του σετ δεδομένων σε μη-επικαλυπτόμενα υποσύνολα εκπαίδευσης, επικύρωσης και ελέγχου με ποσοστά $D_{trn} \rightarrow 60\%$, $D_{val} \rightarrow 20\%$ και $D_{chk} \rightarrow 20\%$ αντιστοίχως. Σύμφωνα με την εκφώνηση, εκπαιδεύονται 4 TSK μοντέλα με διαφορετικές παραμέτρους, στα οποία μεταβάλλονται το πλήθος των συναρτήσεων συμμετοχής και η μορφή των εξόδων όπως φαίνονται παρακάτω στον Πίνακα 1.

	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.

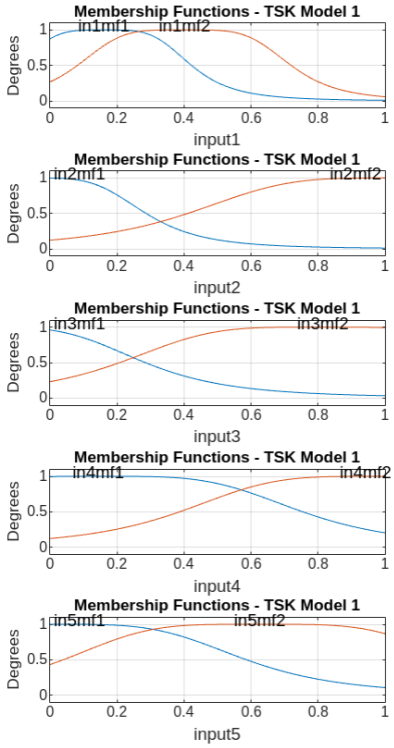
Ο διαχωρισμός και η κανονικοποίση των δεδομένων σε training, validation και check/testing υλοποιήθηκε με την συνάρτηση `split_scale(data, preproc)` που βρίσκεται στο αρχείο `split_scale.m`

Η εκπαίδευση των παραπάνω μοντέλων έγινε σε 200 εποχές.

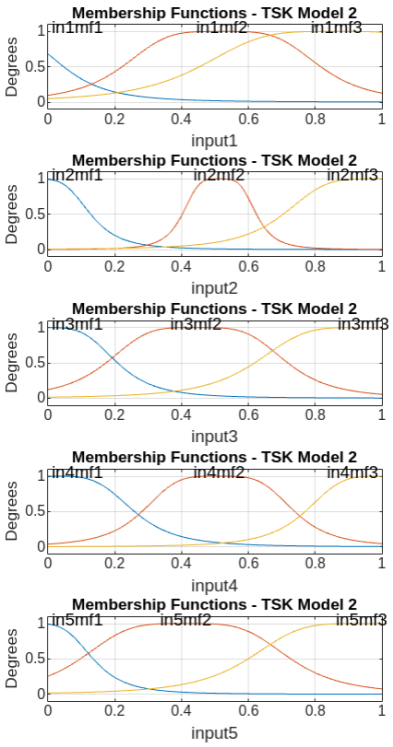
Ως τελικό μοντέλο επιλέχθηκε εκείνο το οποίο αντιστοιχεί στο μικρότερο σφάλμα στο σύνολο validation.

Συναρτήσεις Συμμετοχής

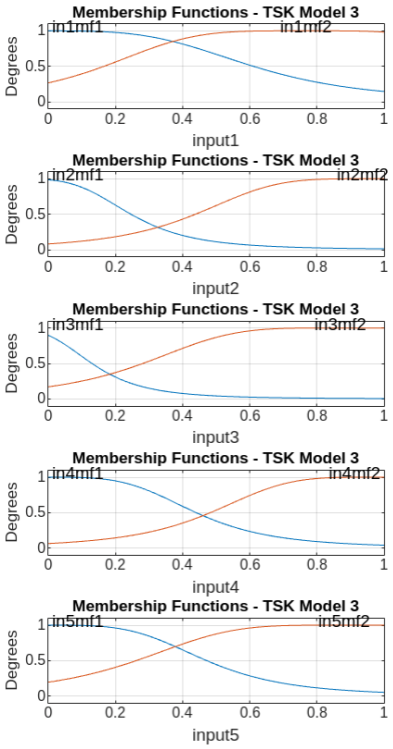
Οι τελικές μορφές των ασαφών συνόλων όπως προέκυψαν από την διαδικασία εκπαίδευσης λήφθηκαν σε διαγράμματα για τις 5 εισόδους των μοντέλων. Αυτές είναι bell-shaped και τα διαδοχικά σύνολα παρουσιάζουν βαθμό επικάλυψης 0.5.



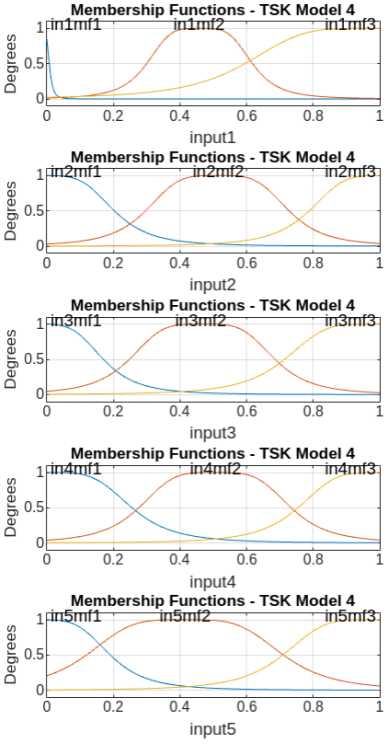
Εικ. 1: Συναρτήσεις συμμετοχής 1ου μοντέλου



Εικ. 2: Συναρτήσεις συμμετοχής 2ου μοντέλου



Εικ. 3: Συναρτήσεις συμμετοχής 3ου μοντέλου

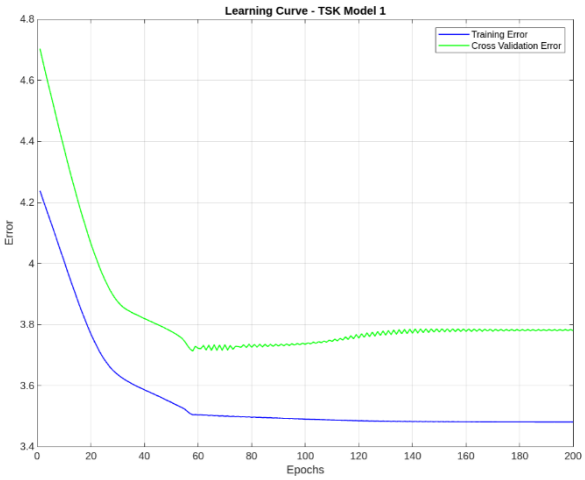


Εικ. 4: Συναρτήσεις συμμετοχής 4ου μοντέλου

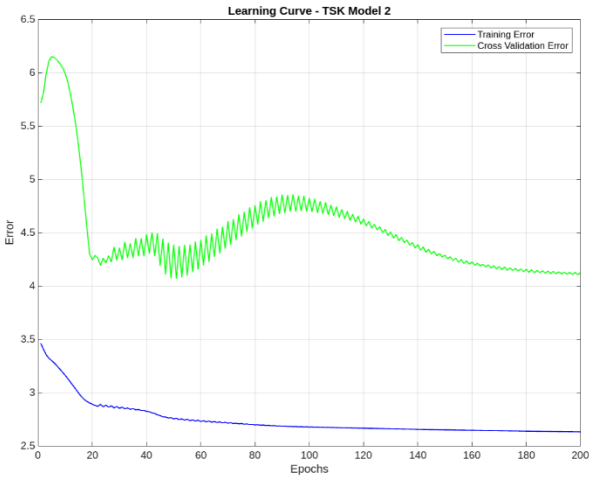
Οι συναρτήσεις συμμετοχής παίζουν σημαντικό ρόλο στη αναλυση που εξετάζουμε καθώς ορίζουν τον τρόπο με τον οποίο κάθε σημείο του χώρου εισόδου αντιστοιχίζεται σε ένα βαθμό μέλους μεταξύ 0 και 1. Υποδεικνύει το βαθμό στον οποίο μια συγκεγκριμένη είσοδος ανήκει σε ένα ασαφές σύνολο και χειρίζεται την αβεβαιότητα του συνόλου δεδομένων.

Διαγράμματα Μάθησης

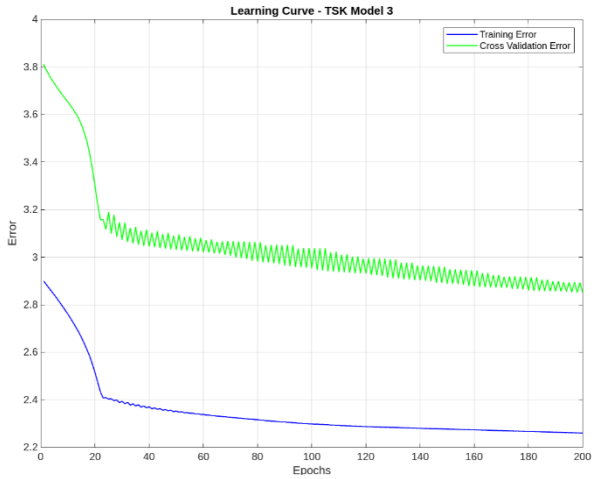
Η απεικόνιση του σφάλματος συναρτήσει του αριθμού επαναλήψεων γίνεται στις καμπύλες μάθησης όπως φαίνεται παρακάτω. Με μπλε χρώμα παρουσιάζεται το σφάλμα εκπαίδευσης και με πράσινο χρώμα το σφάλμα επικύρωσης.



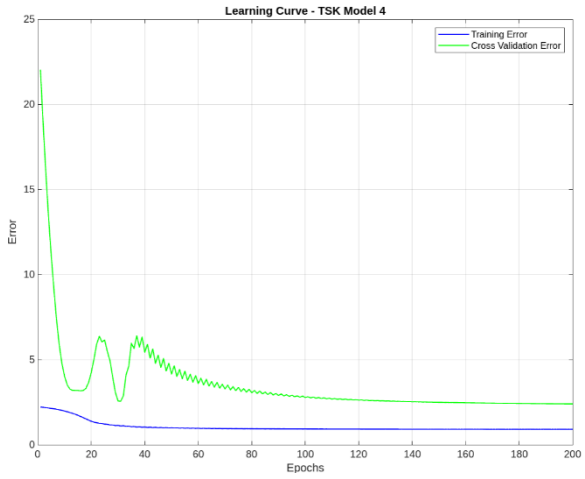
Εικ. 5: Καμπύλες μάθησης 1ου μοντέλου



Εικ. 6: Καμπύλες μάθησης 2ου μοντέλου



Εικ. 7: Καμπύλες μάθησης 3ου μοντέλου



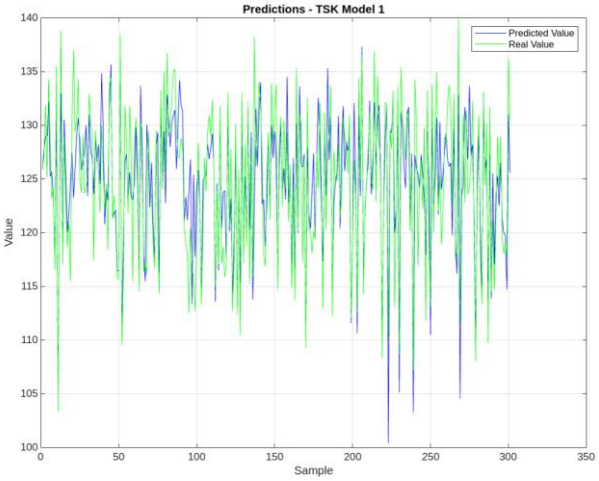
Εικ. 8: Καμπύλες μάθησης 4ου μοντέλου

Παρατηρούμε ότι το σφάλμα εκπαίδευσης σε κάθε μοντέλο, συνεχώς μειώνεται ασυμπτωτικά μέχρι το τέλος της διαδικασίας. Στο 1ο διάγραμμα, η καμπύλη για το σφάλμα επικύρωσης μειώνεται μέχρι τις 60 εποχές και μετά ακολουθεί σταθερή αύξηση θεωρώντας ότι μπαινει στη ζώνη του overfitting. Στο 2ο μοντέλο, λαμβάνουμε αστάθεια και θόρυβο στη καμπύλη για τις πρώτες 90 εποχές ενώ μετέπειτα παρουσιάζεται σταθερή μείωση. Το validation error εδώ απέχει αρκετά από τις ιδανικές τιμές.

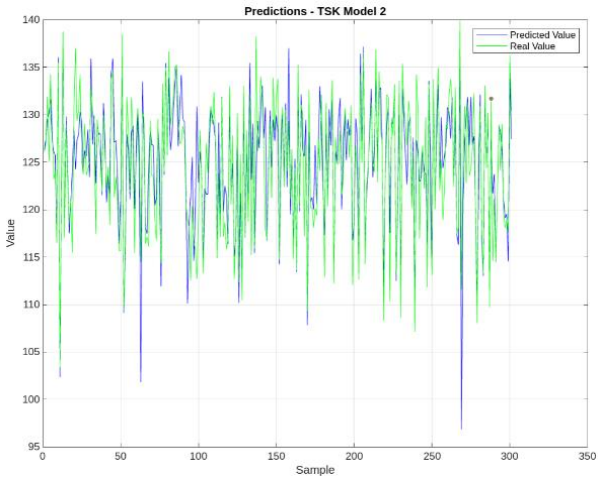
Στο 3ο μοντέλο λαμβάνουμε γρήγορη και σταθερή μείωση του σφάλματος σε όλο τη περίοδο που εξετάζουμε. Το σφάλμα επικύρωσης παρουσιάζει θόρυβο, βρισκεται μακριά από το σφάλμα εκπαίδευσης αλλά παρακολουθούμε μια σταθερή προσαρμογή στη διαδικασία. Εμφανώς καλύτερα αποτελέσματα παίρνουμε στο 4ο μοντέλο, παρά τον θόρυβο αρχικά, όπου τόσο τα σφάλματα επικύρωσης όσο και τα σφάλματα εκπαίδευσης είναι χαμηλά και κοντά το ένα στο άλλο.

Σφάλματα Πρόβλεψης

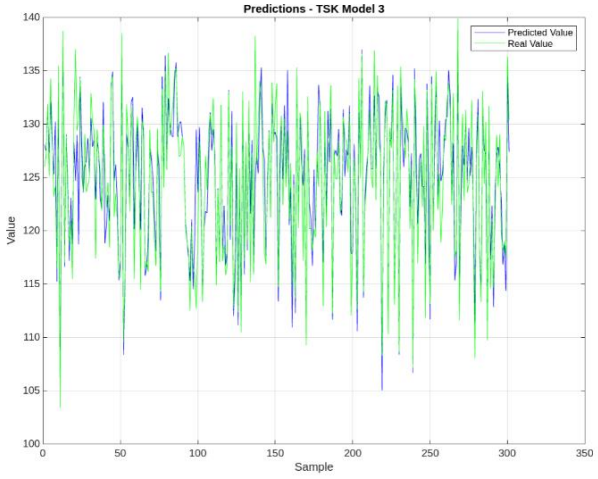
Τα σφάλματα πρόβλεψης μας βοηθούν να αποτυπώσουμε τη σύγκριση ανάμεσα στη προβλεπώμενη και στη πραγματική τιμή της εξόδου στο σύνολο ελέγχου. Με πράσινο χρώμα δίνεται η πραγματική τιμή, ενώ με μπλε η πρόβλεψη του μοντέλου.



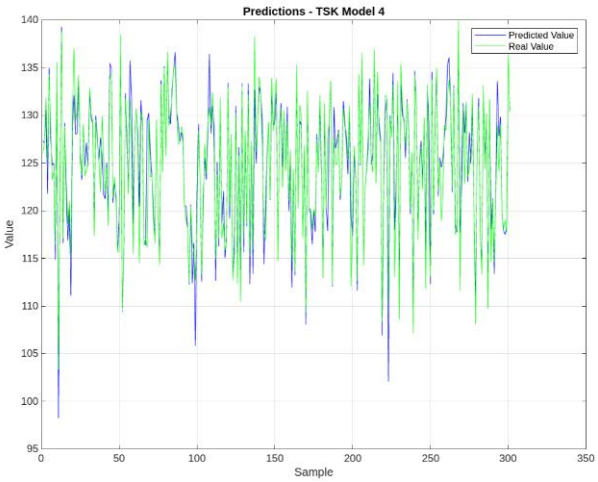
Εικ. 9: Σφάλματα πρόβλεψης 1ου μοντέλου



Εικ. 10: Σφάλματα πρόβλεψης 2ου μοντέλου



Εικ. 11: Σφάλματα πρόβλεψης 3ου μοντέλου



Εικ. 12: Σφάλματα πρόβλεψης 4ου μοντέλου

Παρατηρούμε ότι στο 1ο και 2ο μοντέλο παλινδρόμησης υπάρχουν ορισμένα λάθη στη πρόβλεψη των αποτελεσμάτων. Αυτά εντοπίζονται ομοίως για τα δύο μοντέλα γυρω από την περιοχή με δείγμα (sample)→100 και τιμή (value) →130, όπου η εκτίμηση αστοχεί να παρακολουθήσει τα δείγματα του συνόλου ελέγχου. Στο 1ο μοντέλο, η περιοχή με sample→ 225 και value→ 105 λαμβάνει πολύ μικρότερες τιμές στην εκτίμηση μας έναντι της πραγματικής τιμης, κατι το οποίο συμβαίνει και στην περιοχή με sample→ 275 του 2ου μοντέλου. Αντίθετα, τα μοντέλα 3 και 4

φαίνεται να έχουν καλύτερη προσαρμογή στις αντιστοιχες περιοχές και αποτυπώνουν καλά την υποκείμενη σχέση των δεδομένων.

Δείκτες Απόδοσης

Οι μετρικές που υπολογίστηκαν με σκοπό την ακρίβεια της εκτίμησης της πραγματικής συνάρτησης σε καθένα από τα τέσσερα μοντέλα είναι οι $RMSE, NMSE, NDEI, R^2$. Παρουσιάζονται ακολούθως στον παρακάτω πίνακα.

Μοντέλο	Πλήθος συναρτήσεων συμμετοχής	Μορφή εξόδου	$RMSE$	$NMSE$	$NDEI$	R^2
TSK_model_1	2	Singleton	**3.8498**	**0.2902**	**0.5387**	**0.7097**
TSK_model_2	3	Singleton	**3.6710**	**0.2640**	**0.5138**	**0.7359**
TSK_model_3	2	Polynomial	**2.9877**	**0.1748**	**0.4181**	**0.8252**
TSK_model_4	3	Polynomial	**2.1954**	**0.0943**	**0.3072**	**0.9056**

Πίνακας 2: Μετρικές εκτίμησης σφάλματος.

Συμπεράσματα

Η εκπαίδευση του 4ου μοντέλου διήρκησε περισσότερο από τις άλλες περιπτώσεις διότι είχαμε 1503 παραμέτρους συγκριτικά με το 2ο μοντέλο που λάβαμε συνολικά 288. Αντίστοιχη σύγκριση παρατηρείται μεταξύ 1ου – 3ου μοντέλου όπου οι παράμετροι της ανάλυσης είναι 62 έναντι 222.

Σύμφωνα με το πλήθος συναρτήσεων συμμετοχής

Συγκρίνοντας τα μοντέλα 1-2 που έχουν έξοδο Singleton, παρατηρούμε ότι οι δείκτες απόδοσης λαμβάνουν καλύτερες παραμέτρους στη δεύτερη περίπτωση. Αυτό συμβαίνει διότι το 2ο μοντέλο αριθμεί επιπλέον μια συνάρτηση συμμετοχής, δηλαδή το σύστημα κατέχει περισσότερους κόμβους και περισσότερους ασαφείς κανόνες. Ο συντελεστής προσδιορισμού είναι μεγαλύτερος στο 2ο μοντέλο. Μια υψηλότερη τιμή του R^2 είναι καλύτερη επειδή υποδεικνύει ότι το μοντέλο εξηγεί ένα μεγαλύτερο μέρος της μεταβλητότητας στα δεδομένα.

Συγκρίνοντας τα μοντέλα 3-4 που έχουν έξοδο Polynomial και την ίδια διαφορά ανάμεσα στα ασαφή τους σύνολα, βλέπουμε ότι οι εκτιμήσεις είναι παρόμοιες. Ομοίως οι τρεις πρώτοι δείκτες παρουσιάζουν χαμηλότερες τιμές για το 4ο μοντέλο με τη μετρική R^2 να είναι μεγαλύτερη. Θεωρούμε επομένως αποδοτικότερο το μοντέλο 4 το οποίο σύμφωνα με τα προηγούμενα διαγράμματα δείχνει να γενικεύεται καλύτερα σε νέα δεδομενα.

Σύμφωνα με τη μορφή εξόδου

Παρατηρώντας τα μοντέλα 1-3 που αριθμούν δύο συναρτήσεις συμμετοχής, συμπεραίνουμε ότι οι αποκλίσεις των μετρικών ειναι πολύ μεγαλύτερες συγκριτικά με τις αποκλίσεις στα μοντέλα 1-2. Αυτό σημαίνει ότι ανάμεσα στα μοντέλα με 2 ασαφή σύνολα εισόδων, βέλτιστο είναι εκείνο με την πολυωνυμική έξοδο. Όμοια συμπεράσματα λαμβάνουμε και στη σύγκριση μεταξύ του 2ου – 4ου μοντέλου.

Μέρος Β : Εφαρμογή σε υψηλής διαστασιμότητας dataset

Δεδομένα και Ζητούμενα

Το δεύτερο μέρος της εργασίας δέχεται το *Superconductivity dataset* από το *UCI repository* με 21263 δείγματα και 81 χαρακτηριστικά. Γίνεται διαχωρισμός του σετ δεδομένων σε μη-επικαλυπτόμενα υποσύνολα εκπαίδευσης, επικύρωσης και ελέγχου με ποσοστά $D_{trn} \rightarrow 60\%$, $D_{val} \rightarrow 20\%$ και $D_{chk} \rightarrow 20\%$ αντίστοιχα. Σύμφωνα με την εκφώνηση, το μέγεθος του dataset καθιστά απαγορευτική μια απλή εφαρμογή ενός TSK μοντέλου, όπως στο πρώτο μέρος της εργασίας. Με σκοπό την ελάττωση της πολυπλοκότητας του συστήματος χρησιμοποιούμε τις μεθόδους επιλογής χαρακτηριστικών και της διαμέρισης διασκορπισμού. Εισάγουμε ακολούθως δύο ελεύθερες παραμέτρους για τον αριθμό των χαρακτηριστικών και για τον αριθμό των ομάδων που δημιουργούνται. Τέλος, επιλέγουμε τις βέλτιστες παραμέτρους σύμφωνα με τη μέθοδο αναζήτησης πλέγματος (grid partitioning).

Μέθοδος Αναζήτησης Πλέγματος

Αφού δημιουργήσουμε το 2-διάστατο πλέγμα, ακολουθεί η μέθοδος της 5-πτυχης διασταυρωμένης επικύρωσης όπου εκπαιδεύουμε το μοντέλο για 30 εποχές και αποθηκεύεται το μέσο σφάλμα σε κάθε επανάληψη. Ομαδοποιούμε τα δεδομένα με τον αλγόριθμο Subtractive Clustering και η επιλογή χαρακτηριστικών γίνεται μεσω της relief συνάρτησης επιλέγοντας τους 20 πλησιέστερους γείτονες του cluster.

Οι ελεύθερες μεταβλητές που επιλέξαμε έχουν τις παρακάτω τιμές:

- $αριθμός\ χαρακτηριστικών\ (characteristics) = [3\ 6\ 9\ 12]$
- $ακτίνα\ ομάδων\ (cluster\ radius) = [0.2\ 0.4\ 0.6\ 0.8\ 1]$

Τα αποτελέσματα των μικρότερων σφαλμάτων RMSE μετά τη σύγκριση και για κάθε συνδυασμό παρουσιάζονται ακολούθως στον παρακάτω πίνακα.

radius characteristics	0.2	0.4	0.6	0.8	1
3	22.8152	23.7869	24.0261	24.7744	26.6104
6	19.6445	21.2756	22.0301	23.5568	23.5817
9	19.0679	20.0759	20.7545	22.0115	22.0122
12	**18.2928**	19.0856	19.9671	21.8822	21.8734

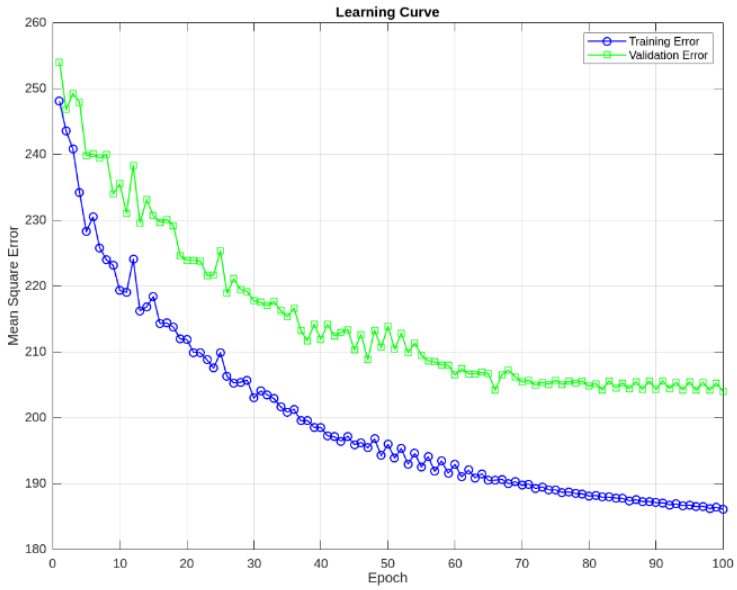
Πίνακας 3: Μετρικές εκτίμησης σφάλματος για τις διάφορες ελεύθερες μεταβλητές.

Παρατηρούμε επομένως ότι το αποδοτικότερο μοντέλο είναι αυτό για ακτίνα ομάδας 0.2 και 12 χαρακτηριστικά.

Βέλτιστο Μοντέλο

Διαγράμμα Μάθησης

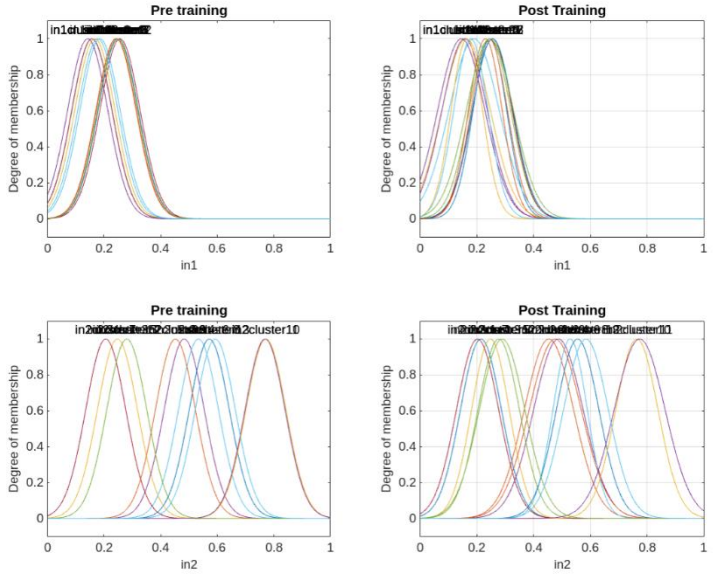
Παραθέτουμε το διάγραμμα μάθησης για το βέλτιστο μοντέλο, το οποίο εκπαιδεύσαμε για 100 εποχές. Με μπλε χρώμα παρουσιάζεται το σφάλμα εκπαίδευσης και με πράσινο χρώμα το σφάλμα επικύρωσης. Τα αποτελέσματα που παίρνουμε από την καμπύλη είναι σωστά και βλέπουμε επιτυχή προσαρμογή στη διαδικασία, εφόσον το validation error μειώνεται σταθερά.



Εικ. 13: Καμπύλη μάθησης βέλτιστου μοντέλου με cluster radius = 0.2 και characteristics = 12

Συναρτήσεις Συμμετοχής

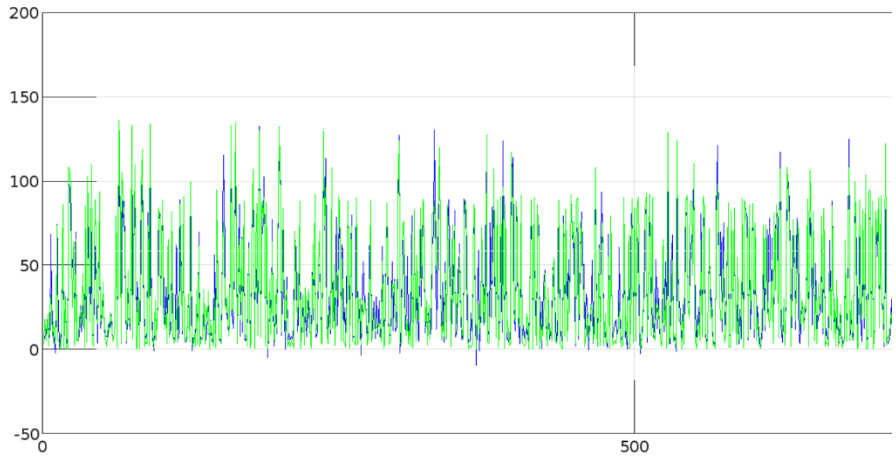
Δίνονται τα ασαφή σύνολα στην αρχική και την τελική τους μορφή μετά τη διαδικασία βελτιστοποίησης. Υπάρχουν 13 συναρτήσεις συμμετοχής με gaussian κατανομή όπως βλέπουμε για τις δύο από τις δώδεκα εισόδους.



Εικ. 14: Συναρτήσεις συμμετοχής του μοντέλου που εξετάζουμε αρχικα και τελικά.

Σφάλματα Πρόβλεψης

Παρουσιάζουμε παρακάτω ενα δείγμα του διαγράμματος για το σφάλμα πρόβλεψης. Σε γενικές γραμμές οι προβλεπώμενες τιμές του μοντέλου προσεγγίζουν τις πραγματικές τιμές του συνόλου επικύρωσης.



Εικ. 15: Διάγραμμα για το σφάλμα πρόβλεψης του μοντέλου.

Δείκτες Απόδοσης

Οι μετρικές που εξάγαμε για να διαπιστώσουμε την ακρίβεια της πραγματικής συνάρτησης είναι οι $RMSE, NMSE, NDEI, R^2$.

$RMSE$	$NMSE$	$NDEI$	R^2
14.2983	**0.172**	**0.41472**	**0.828**

Πίνακας 4: Δείκτες απόδοσης για το μοντέλο επιλογής.

Συμπεράσματα

Είναι εμφανές ότι οι δείκτες NMSE, NDEI, R^2 λαμβάνουν πολύ μικρές και ικανοποιητικές τιμές ενώ μπορούν να συγκριθούν με τα σφάλματα του 1ου μέρους της εργασίας οπου εξετάζαμε ένα μικρο σύνολο δεδομένων. Ωστόσο, το RMSE θεωρείται σχετικά μεγάλο σφάλμα. Το μοντέλο αυτό θεωρούμε οτι αποτυγχάνει σε αυτό τον δείκτη διότι δε χρησιμοποιούμε όλες τις μεταβλητές εισόδου αλλά ένα μέρος απο αυτές για την διαδικασία ανάλυσης.

Οι κανόνες που προκύπτουν απο τη διαδικασία ήταν 13, ενώ σε περίπτωση που διαμερίζαμε τον χώρο εισόδου κάθε μεταβλητής με δύο ασαφή σύνολα, θα καταλήγαμε σε 2^{81} κανόνες. Η πολυπλοκότητα θα ήταν μεγάλη και το μοντέλο θα αποτύγχανε.

Η εργασία υλοποιήθηκε στο περιβάλλον του `Matlab` και το πρόγραμμα εκτελείται μεσω του αρχείου `sys a.m` για το 1ο μέρος και μέσω του αρχείου `sys b.m` για το 2ο μέρος της εργασίας. Ο κώδικας ειναι αναρτημένος σε αποθετήριο στο Github.