

Εργασία 3 Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Φοιτητής : Μπεκιάρης Θεοφάνης ΑΕΜ:8200

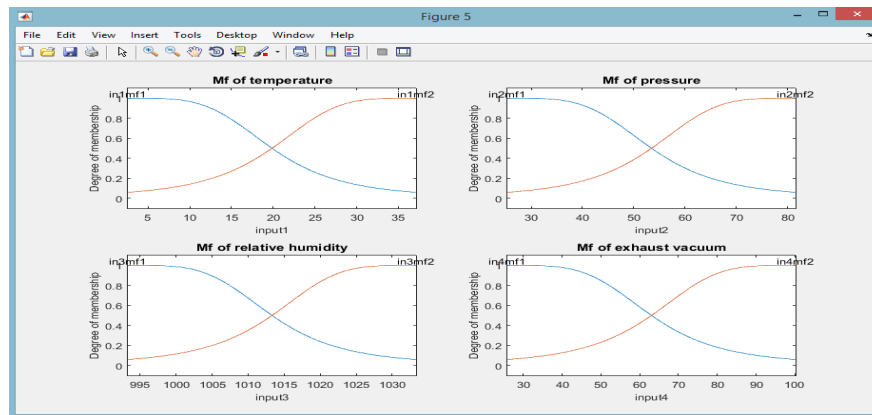
1)Πρώτη εφαρμογή Power Plant dataset(Εκπαίδευση TSK μοντέλων με διαφορετικές παραμέτρους).

Παρακάτω θα παρουσιαστούν τα αποτελέσματα από την εκπαίδευση των 4 TSK μοντέλων τα οποία διαφέρουν ως προς το πλήθος συναρτήσεων συμμετοχής των εισόδων και την μορφή της εξόδου τους. Για την κατασκευή των μοντέλων έχει χρησιμοποιηθεί η συνάρτηση genfis1 του Matlab που δημιουργεί TSK μοντέλο και αρχικοποιεί τις συναρτήσεις συμμετοχής σε μορφή bell-shaped όπως ζητείται στην εκφώνηση και η εκπαίδευση του μοντέλου γίνεται μέσω της συνάρτησης anfis με χρήση της υβριδικής μεθόδου. Τέλος για την διαδικασία εκπαίδευσης,επικύρωση και ελέγχου έχουμε διαχωρίσει τα δεδομένα εκπαίδευσης,επικύρωσης και ελέγχου όπως ζητείται,δηλαδή σε ποσοστό 60%, 20% και 20% έτσι ώστε τα δεδομένα να μην εμφανίζουν επικάλυψη μεταξύ τους.

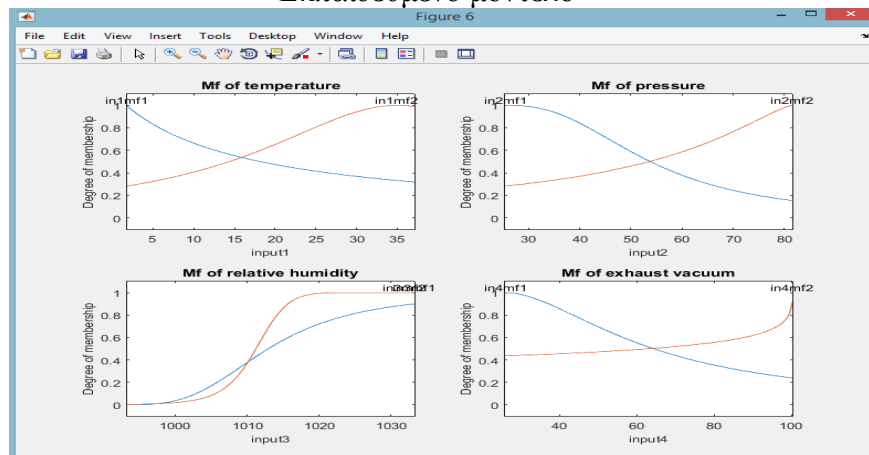
TSK model 1 με 2 συναρτήσεις συμμετοχής και έξοδο Sigleton

Παρακάτω παρουσιάζονται οι μορφές των συναρτήσεων συμμετοχής πριν και μετά την διαδικασία εκπαίδευσης του μοντέλου. Η εκπαίδευση έχει γίνει για 300 εποχές. Τα δεδομένα πριν κάθε διαδικασία εκπαίδευσης ανακατεύονται για να μην έχουν την ίδια σειρά ώστε τα μοντέλα που προκύπτουν κάθε φορά να είναι περισσότερο αξιόπιστα.

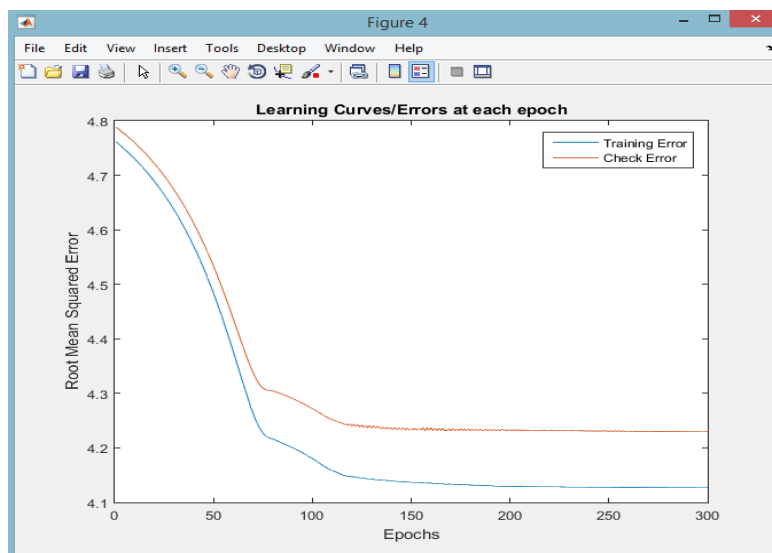
Αρχικό μοντέλο



Εκπαιδευμένο μοντέλο



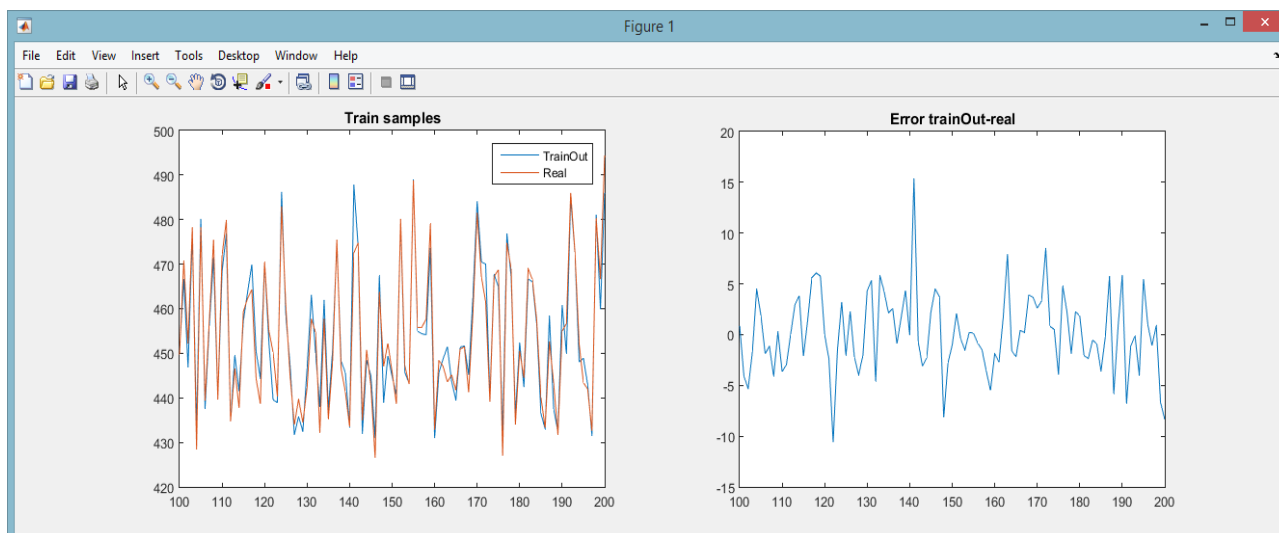
Στην συνέχεια παρατίθενται το διάγραμμα μάθησης(Learning Curves) στο οποίο παρουσιάζεται το σφάλμα του μοντέλου συναρτήσει των επαναλήψεων/εποχών. Η πορτοκαλί καμπύλη αντιστοιχεί στο σφάλμα που δίνει το validation δείγμα δεδομένων(valData) που εισάγουμε στην anfis. Ο σκοπός της χρήσης του validation δείγματος δεδομένων είναι για την αποφυγή υπερ-εκπαίδευσης του μοντέλου. Περισσότερα για το φαινόμενο της υπερεκπαίδευση θα συζητηθούν στην συνέχεια.



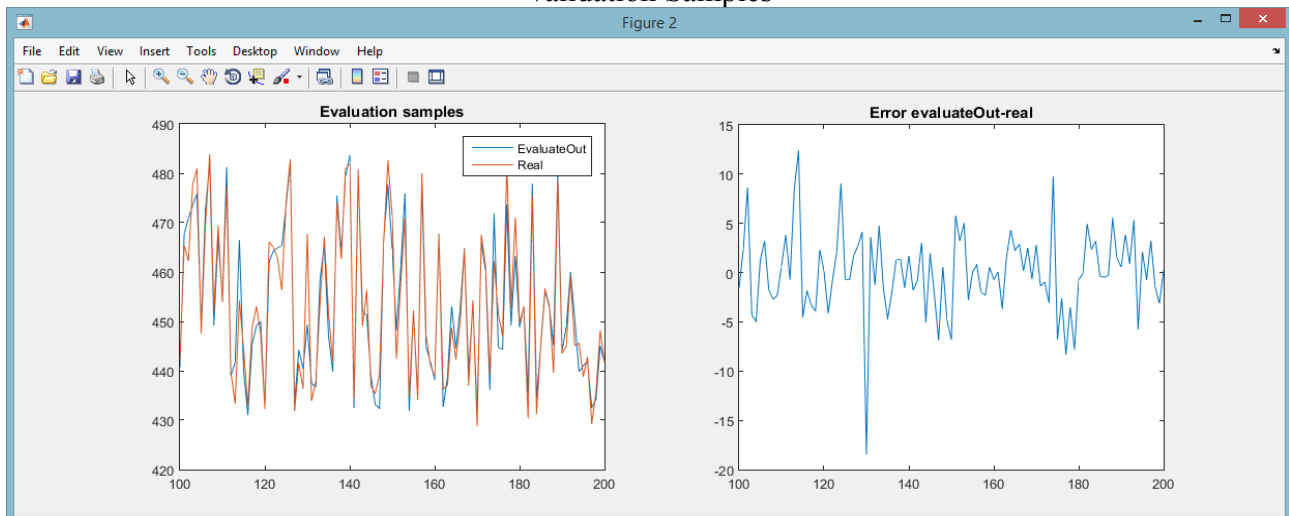
Στο παραπάνω διάγραμμα βλέπουμε ότι μετά από κάποιο σημείο,περίπου στις 150 εποχές το μοντέλο παύει να εκπαιδεύεται γρήγορα και το σφάλμα αποκτάει σχεδόν σταθερή τιμή ή διαφορετικά θα λέγαμε ότι το μοντέλο συνεχίζει να εκπαιδεύεται με πάρα πολύ μικρό ρυθμό,σχεδόν αμελητέο.

Στην συνέχεια παρουσιάζονται και τα διαγράμματα σφάλματος για το τελικό εκπαιδευμένο μοντέλο και τα τρία είδη δεδομένων που διαχωρίσαμε,δηλαδή τα training,validation και check δεδομένα. Για την καλύτερη παρουσίαση των γραφικών παραστάσεων παρουσιάζετε η γραφική παράσταση από για 100 δείγματα από κάθε σύνολο(μεταξύ 100 και 200).

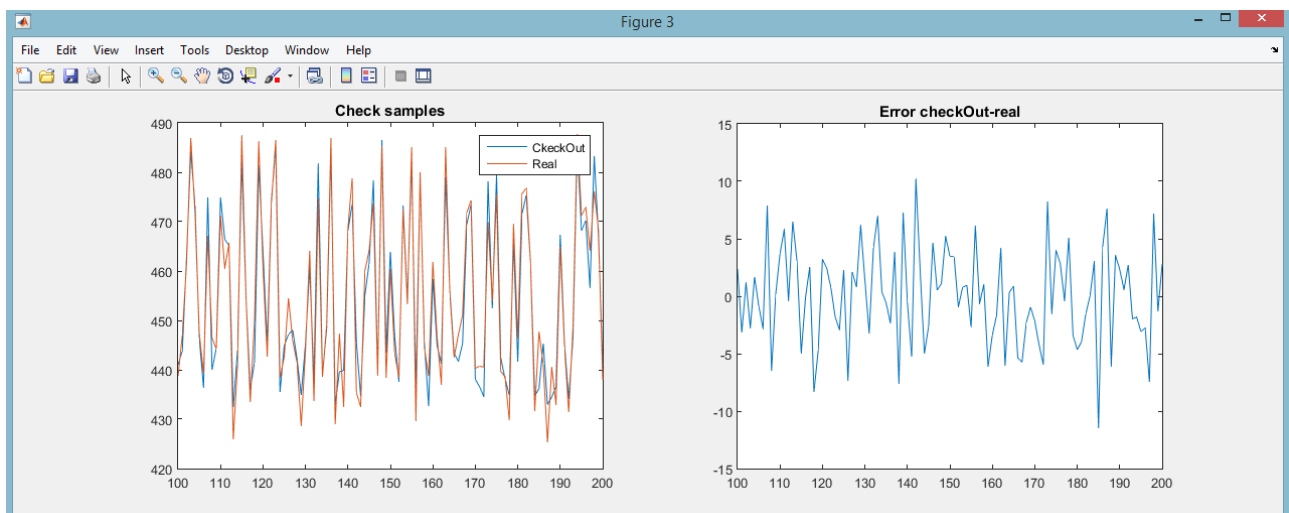
Training Sample



Validation Samples

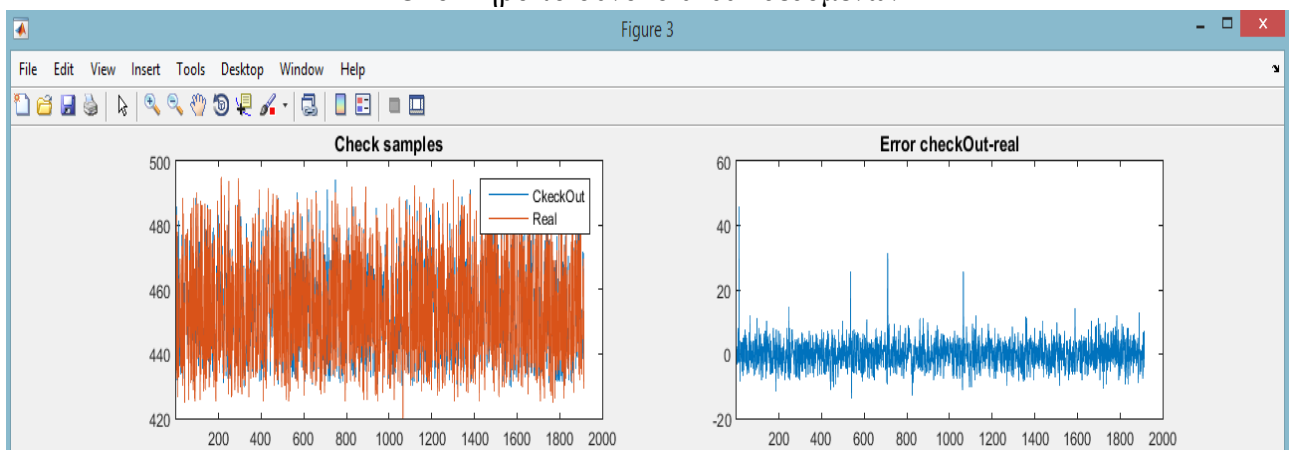


Check samples



Επιπλέον θα παρουσιάσουμε και το διάγραμμα σφάλματος για ολόκληρο το σύνολο check δεδομένων απλά για λόγους επίδειξης της σωστής συμπεριφοράς του μοντέλου πάνω σε ολόκληρο το σύνολο και όχι απλά σε ένα κομμάτι δεδομένων μεταξύ του 100 και 200. Έτσι έχουμε.

Ολόκληρο το σύνολο check δεδομένων

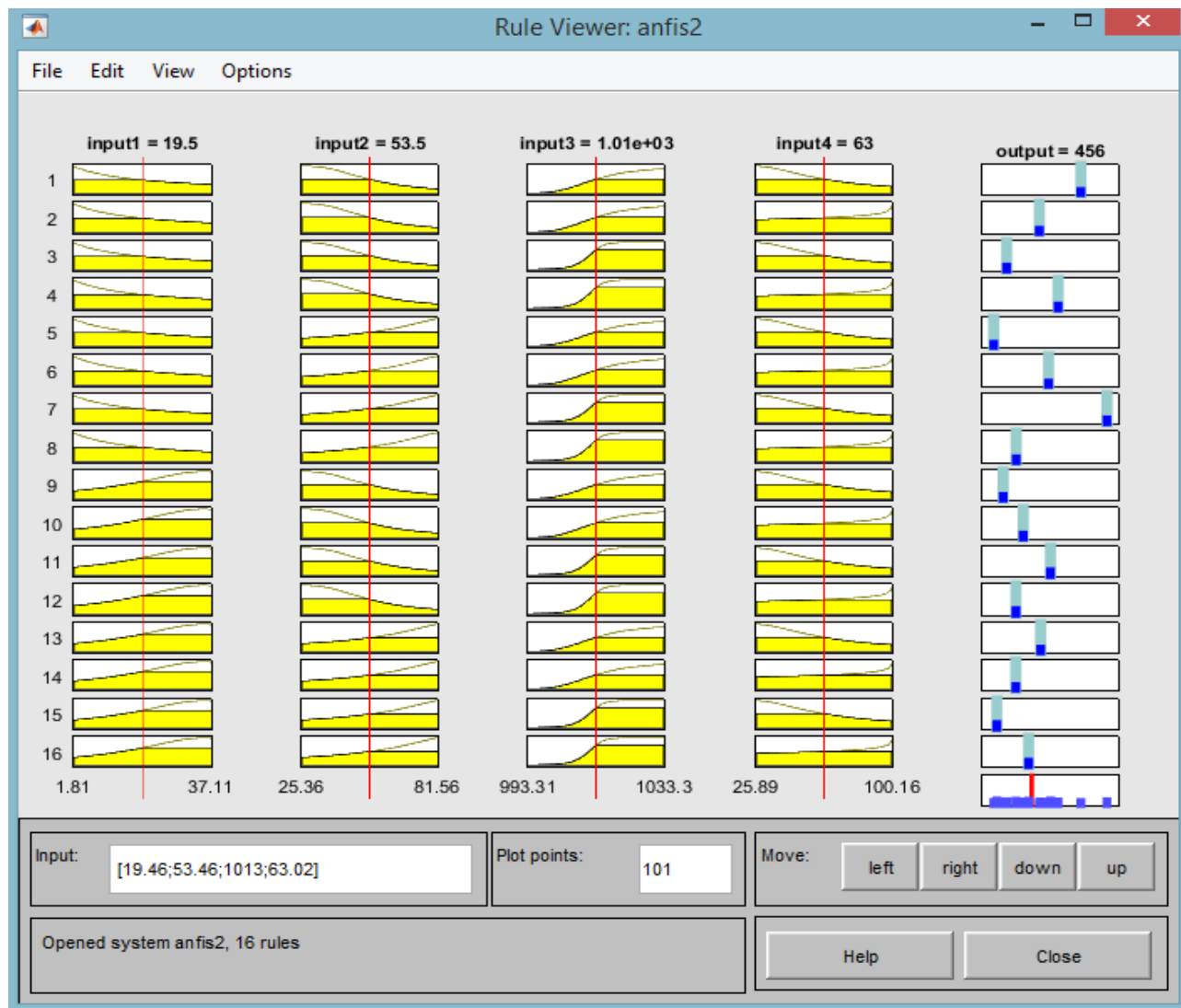


Μπεκιάρης Θεοφάνης ΑΕΜ:8200 Ασαφή Συστήματα 2018

Μεγαλύτερης σημασίας είναι τα αποτελέσματα που προκύπτουν από τα check δεδομένα καθώς αυτά είναι δεδομένα άγνωστα για το μοντέλο και μπορούν να δείξουν πόσο καλά έχει εκπαιδευτεί το μοντέλο και κατά πόσο είναι ικανό να χρησιμοποιηθεί για σωστή πρόβλεψη. Τα training και validation δεδομένα έχουν ξανά χρησιμοποιηθεί από το μοντέλο για την διαδικασία εκπαίδευσης άρα το μοντέλο μπορεί απλά να έχει προσαρμοστεί πάνω σε αυτά τα δεδομένα ώστε να προκύπτει ελάχιστο σφάλμα αλλά να μην έχει εκπαιδευτεί κατάλληλα ώστε να κάνει σωστή πρόβλεψη για οποιαδήποτε δεδομένα του δείγματος δεδομένων. Τα διαγράμματα σφάλματος για τα training και validation δεδομένα παρουσιάζονται για λόγους πληρότητας ώστε να φανεί η συμπεριφορά του μοντέλου πάνω σε ολόκληρο το δείγμα δεδομένων.

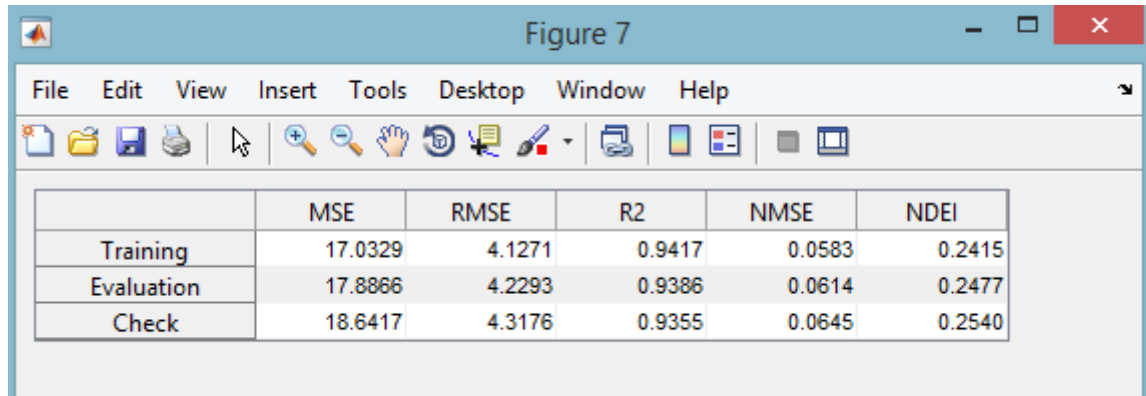
Οι κανόνες που διαμορφώνονται για 4 εισόδους και 2 συναρτήσεις συμμετοχής είναι $2^4 = 16$ κανόνες. Για την παρουσίαση των κανόνων του μοντέλου καλούμε την fuzzy(tuned_fis) και επιλέγουμε από το GUI view rules, έτσι παίρνουμε.

Κανόνες βάσης



Τέλος παρουσιάζουμε τους δείκτες απόδοσης που αναφέρονται στην εκφώνηση. Με χρήση της συνάρτησης `unitable` του Matlab έχουμε τοποθετήσει τους δείκτες σε πίνακα για την καλύτερη παρουσίαση.

Πίνακας δεικτών απόδοσης



	MSE	RMSE	R2	NMSE	NDEI
Training	17.0329	4.1271	0.9417	0.0583	0.2415
Evaluation	17.8866	4.2293	0.9386	0.0614	0.2477
Check	18.6417	4.3176	0.9355	0.0645	0.2540

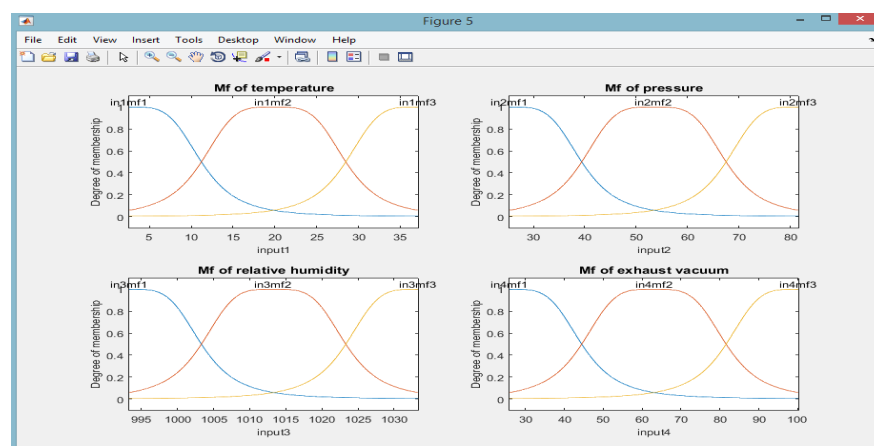
Από τα παραπάνω διαγράμματα βλέπουμε ότι το μοντέλο 1 TSK έχει αρκετά καλή απόδοση όσον αφορά την εκπαίδευση του και θα μπορούσε να χρησιμοποιηθεί για την παραγωγή αξιόπιστων προβλέψεων. Για τα παραπάνω check δεδομένα παίρνουμε συντελεστή προσδιορισμού πολύ κοντά στην μονάδα, $R^2 = 0.9355$ που δείχνει ότι έχει γίνει καλή εκτίμηση. Το μέσο τετραγωνικό σφάλμα που παίρνουμε είναι $MSE = 18.64$ δηλαδή στην τάξη των δεκάδων, δεδομένου όμως ότι οι απόλυτες τιμές των δειγμάτων της εξόδου είναι της τάξης των εκατοντάδων όπως βλέπουμε και από τα διαγράμματα έχουμε τιμές περίπου μεταξύ του 430 και 490, άρα ένα σφάλμα της τάξης του 18 είναι περίπου $18/450 = 0.04$ ή 4% σφάλμα.

Στην συνέχεια θα παρουσιαστούν τα ίδια διαγράμματα και για τα υπόλοιπα μοντέλα με σκοπό την σύγκριση των μοντέλων και την εξαγωγή γενικών συμπερασμάτων για την απόδοση των μοντέλων βάση των χαρακτηριστικών τους.

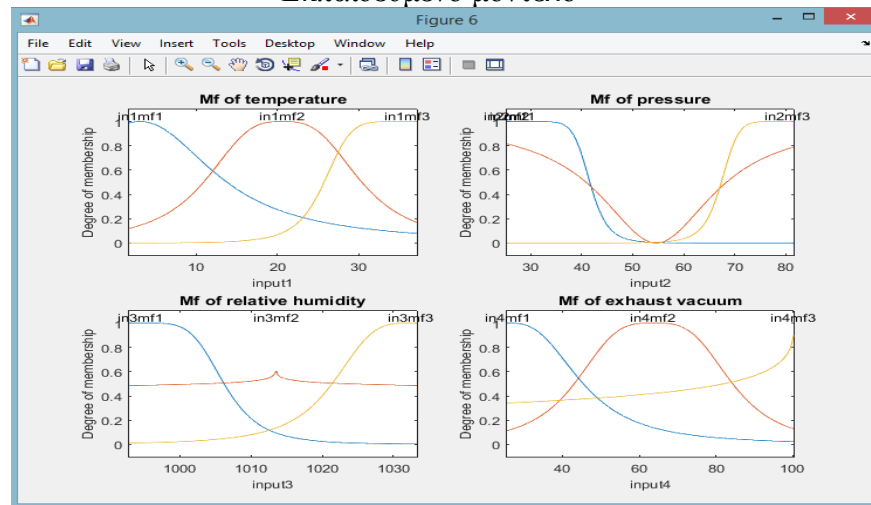
TSK model 2 με 3 συναρτήσεις συμμετοχής και έξοδο Sigleton

Μορφή συναρτήσεων συμμετοχής πριν και μετά την διαδικασία εκπαίδευσης του μοντέλου. Η εκπαίδευση έχει γίνει για 300 εποχές. Τα δεδομένα πριν κάθε διαδικασία εκπαίδευσης ανακατεύονται για να μην έχουν την ίδια σειρά ώστε τα μοντέλα που προκύπτουν κάθε φορά να είναι περισσότερο αξιόπιστα.

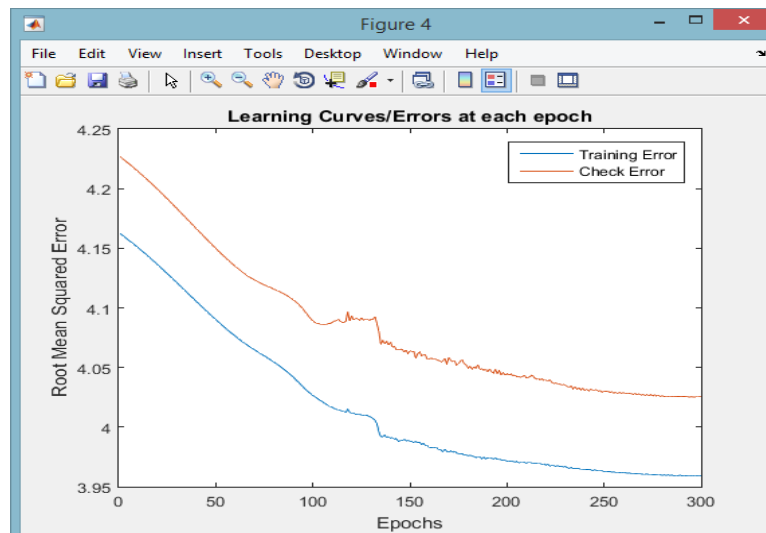
Αρχικό μοντέλο



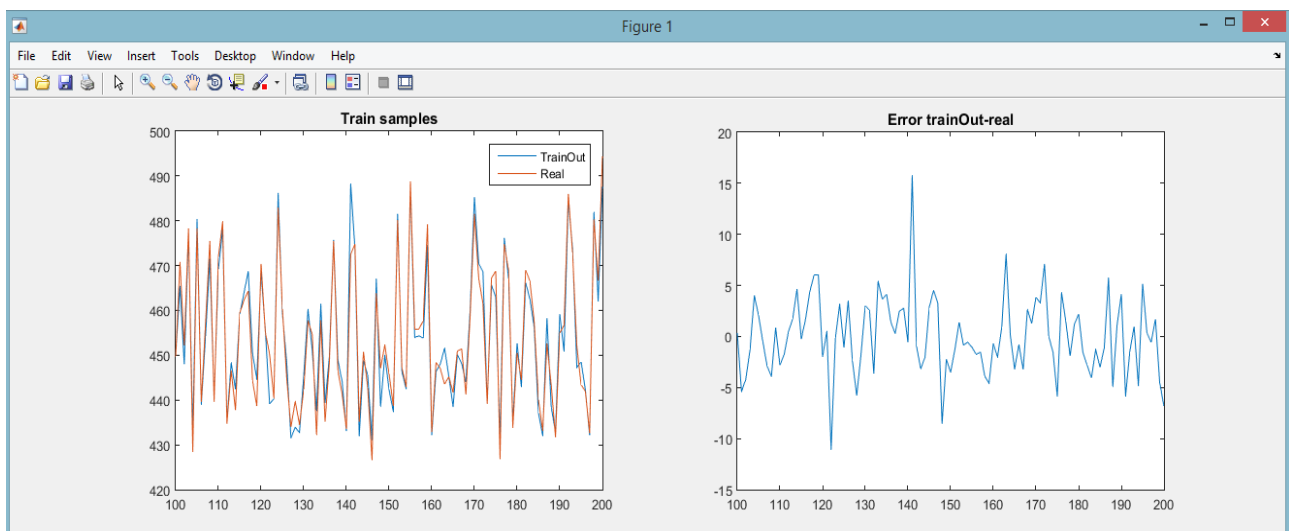
Εκπαιδευμένο μοντέλο



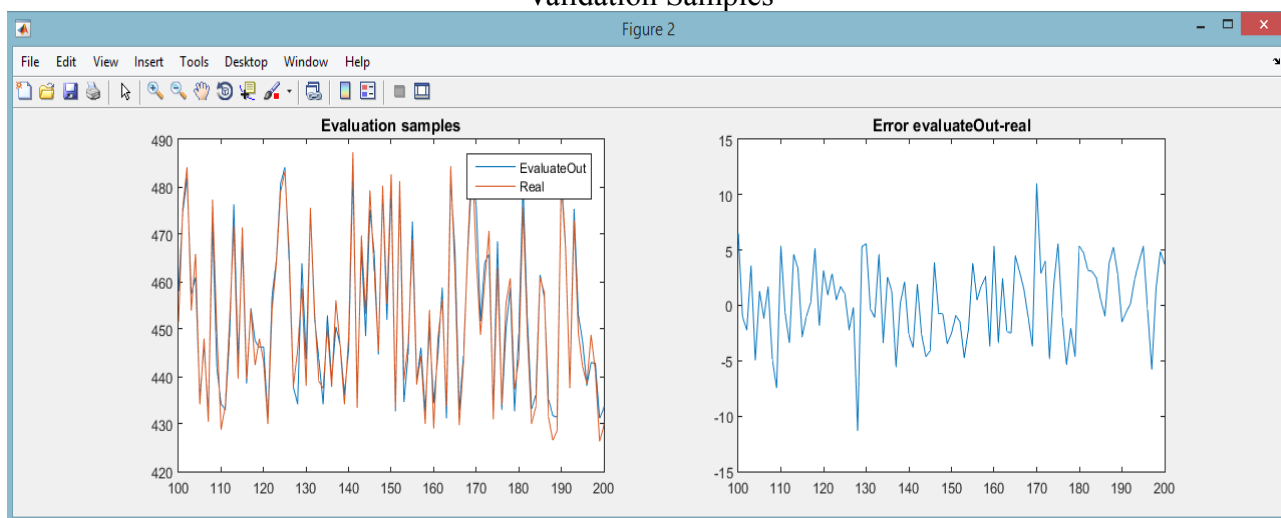
Διάγραμμα μάθησης (Learning Curves)



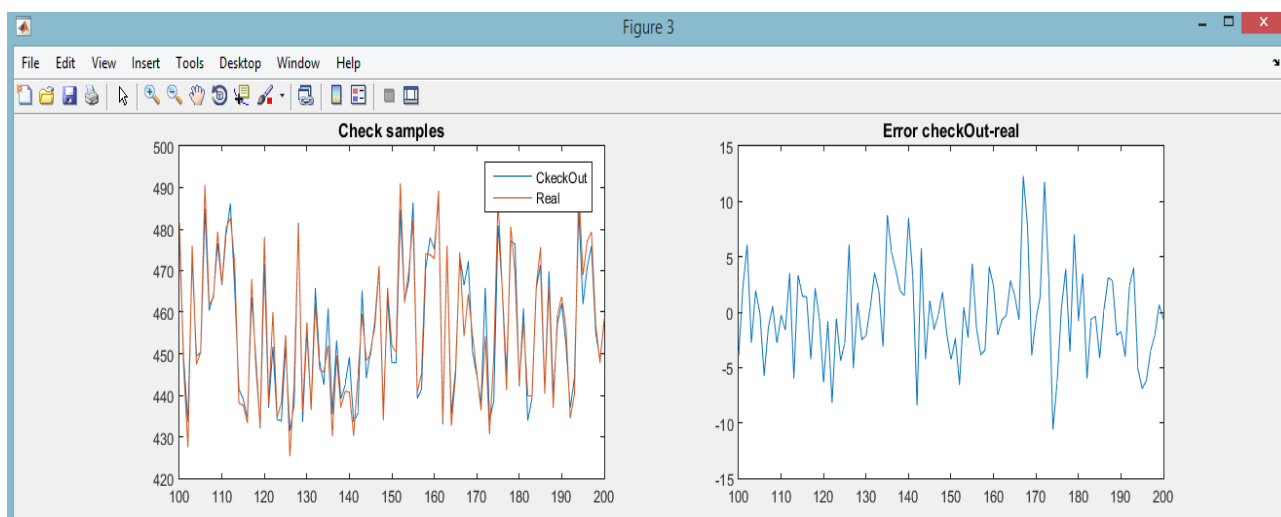
Training Sample



Validation Samples

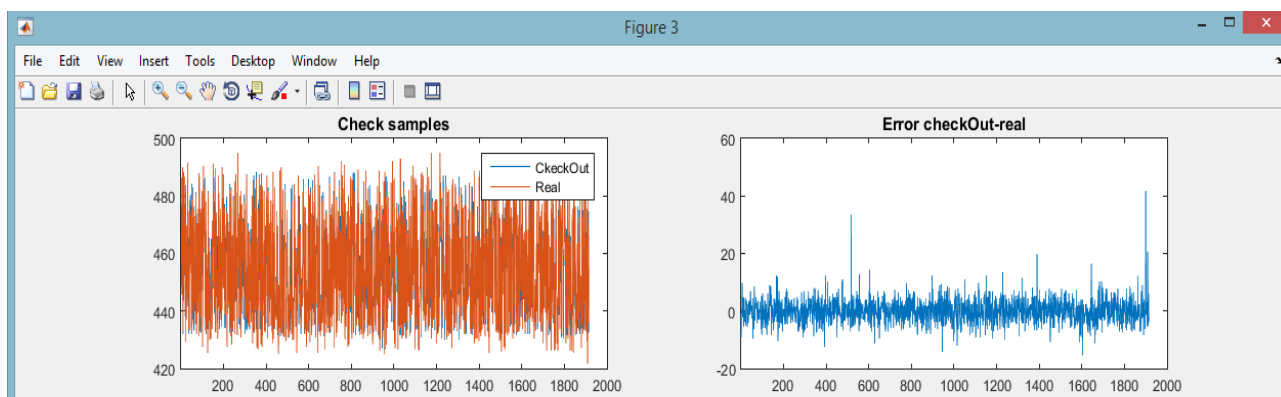


Check samples



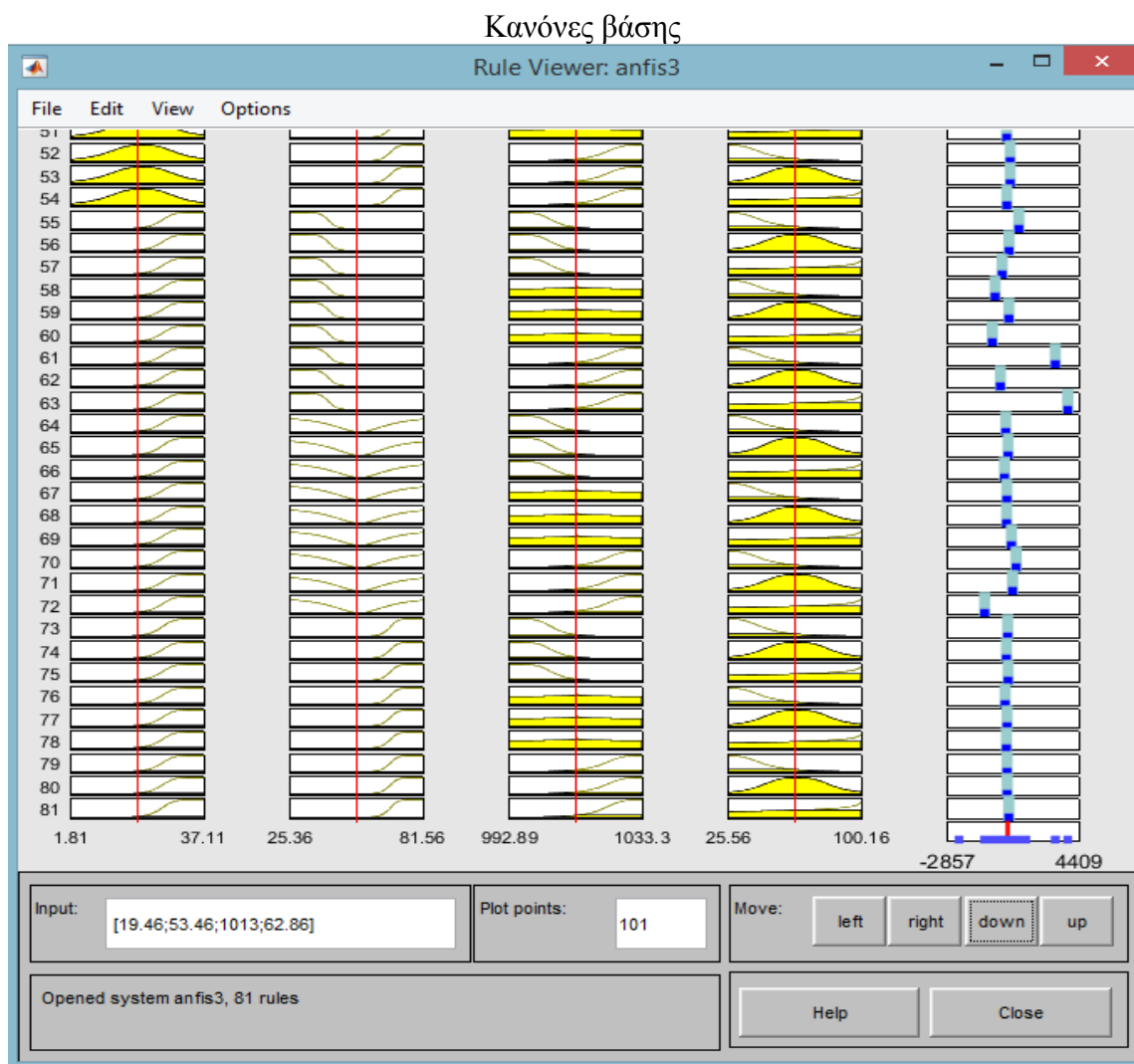
Επιπλέον θα παρουσιάσουμε και το διάγραμμα σφάλματος για ολόκληρο το σύνολο check δεδομένων απλά για λόγους επίδειξης της σωστής συμπεριφοράς του μοντέλου πάνω σε ολόκληρο το σύνολο και όχι απλά σε ένα κομμάτι δεδομένων μεταξύ του 100 και 200. Έτσι έχουμε.

Ολόκληρο το σύνολο check δεδομένων



Όπως φαίνεται στο διάγραμμα το μοντέλο λειτουργεί για όλο το σύνολο δεδομένων.

Οι κανόνες που διαμορφώνονται για 4 εισόδους και 3 συναρτήσεις συμμετοχής είναι $3^4 = 81$ κανόνες. Για την παρουσίαση των κανόνων του μοντέλου καλούμε την fuzzy(tuned_fis) και επιλέγουμε από το GUI view rules.



Πίνακας δεικτών απόδοσης

	MSE	RMSE	R2	NMSE	NDEI
Training	15.6726	3.9589	0.9460	0.0540	0.2324
Evaluation	16.2026	4.0252	0.9438	0.0562	0.2371
Check	16.3491	4.0434	0.9451	0.0549	0.2343

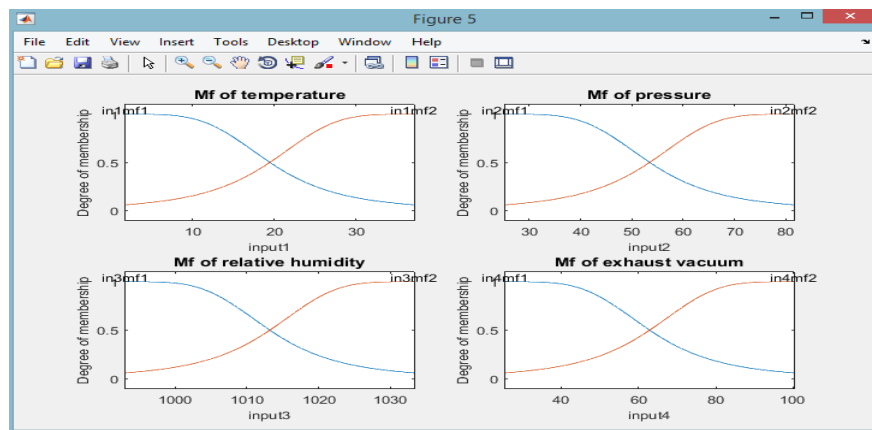
Παρατηρούμε ότι το μοντέλο 2 κάνει καλύτερες προβλέψεις από το μοντέλο 1 καθώς έχει μειωθεί το μέσο τετραγωνικό σφάλμα MSE καθώς και οι τιμές NMSE και NDEI, ενώ ο συντελεστής προσδιορισμού R^2 έχει αυξηθεί.

Στην συνέχεια θα δούμε την συμπεριφορά των μοντέλων με πολωνυμική έξοδο.

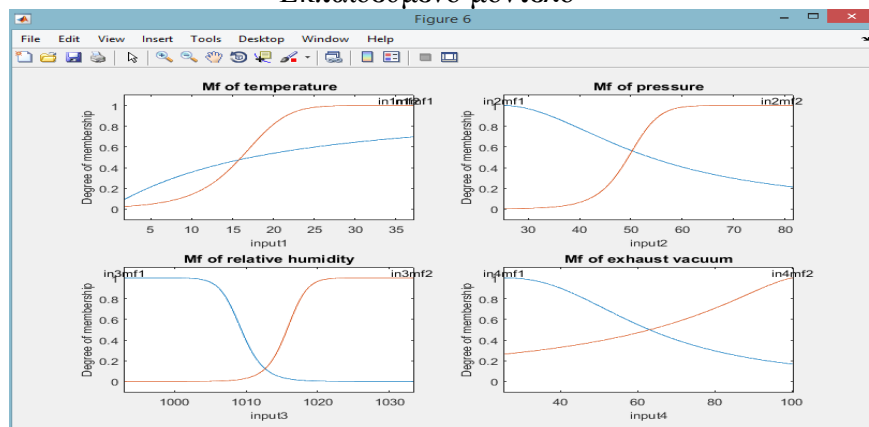
TSK model 3 με 2 συναρτήσεις συμμετοχής και έξοδο Polynomial

Μορφή συναρτήσεων συμμετοχής πριν και μετά την διαδικασία εκπαίδευσης του μοντέλου. Η εκπαίδευση έχει γίνει για 300 εποχές. Τα δεδομένα πριν κάθε διαδικασία εκπαίδευσης ανακατεύονται για να μην έχουν την ίδια σειρά ώστε τα μοντέλα που προκύπτουν κάθε φορά να είναι περισσότερο αξιόπιστα.

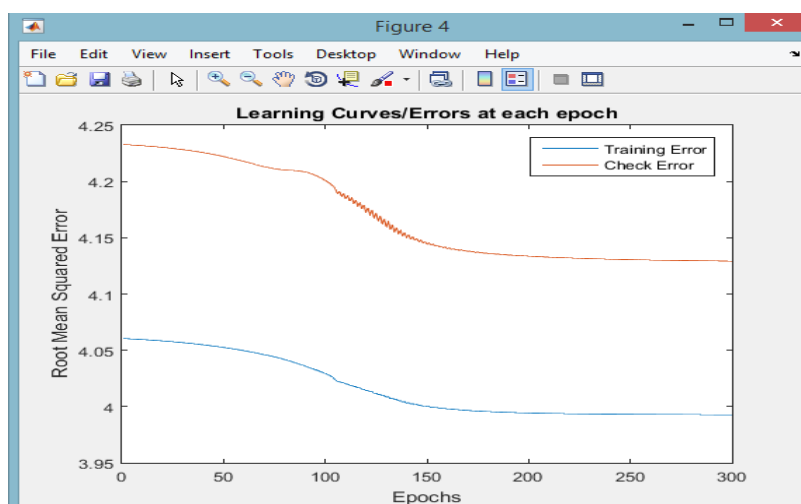
Αρχικό μοντέλο



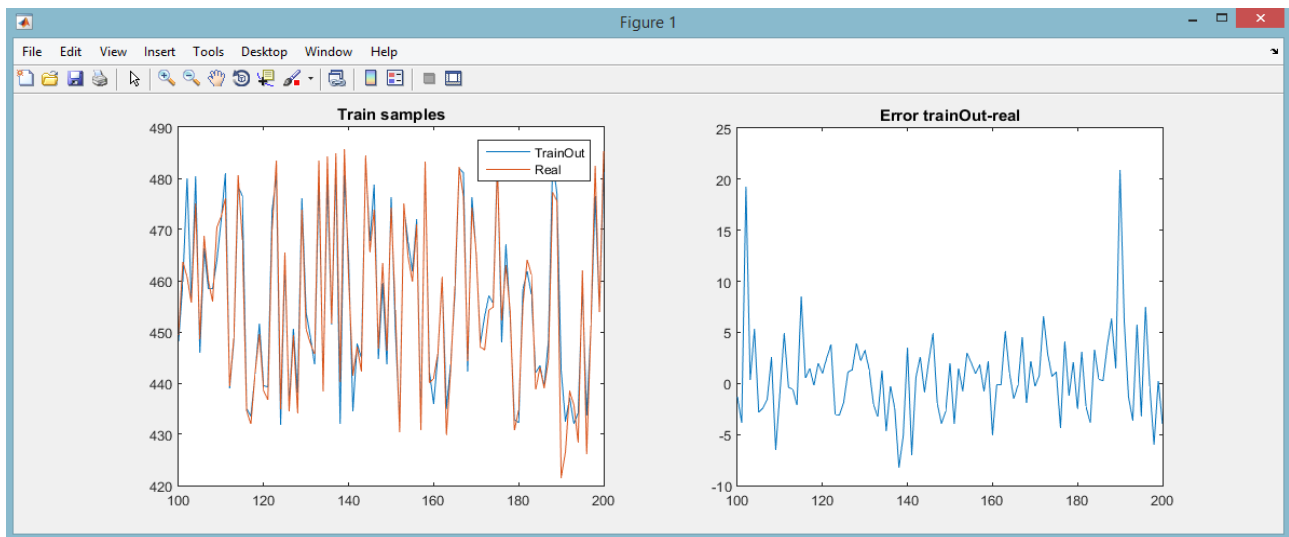
Εκπαιδευμένο μοντέλο



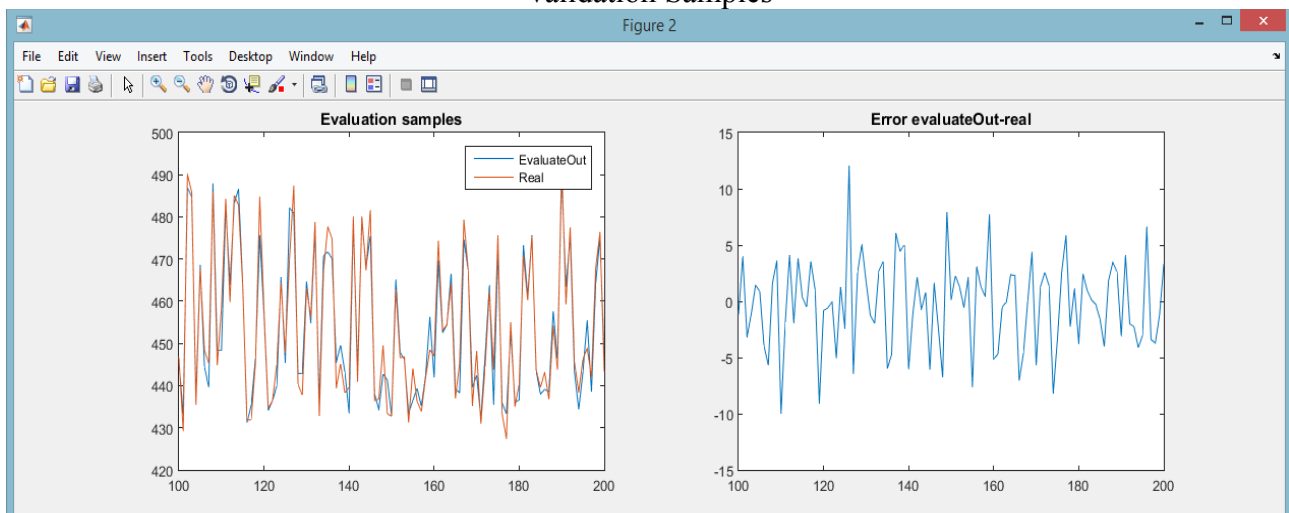
Διάγραμμα μάθησης (Learning Curves)



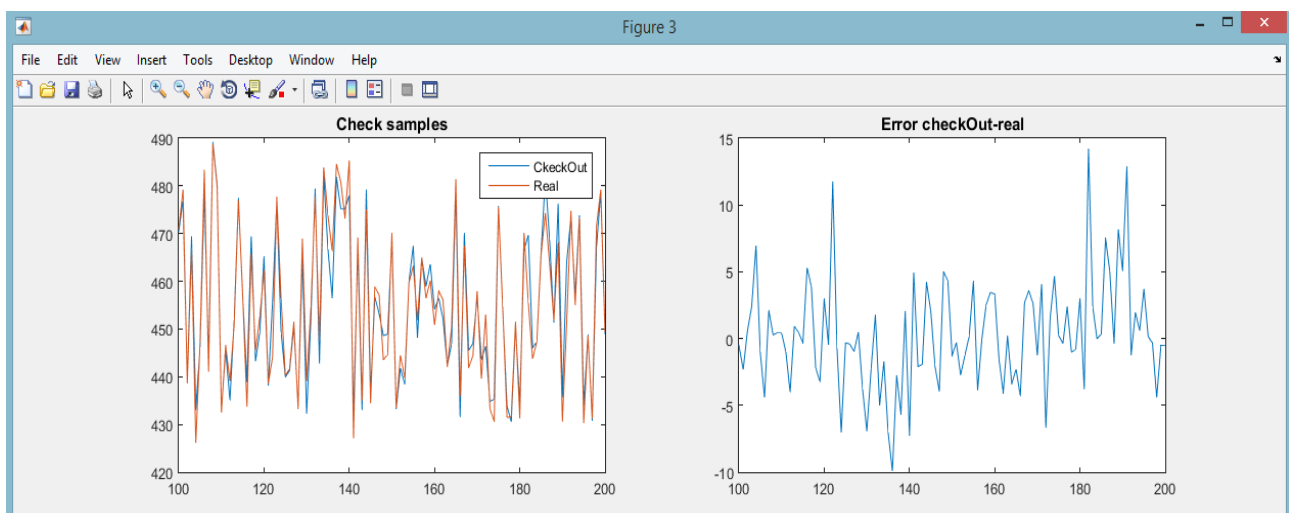
Training Sample



Validation Samples

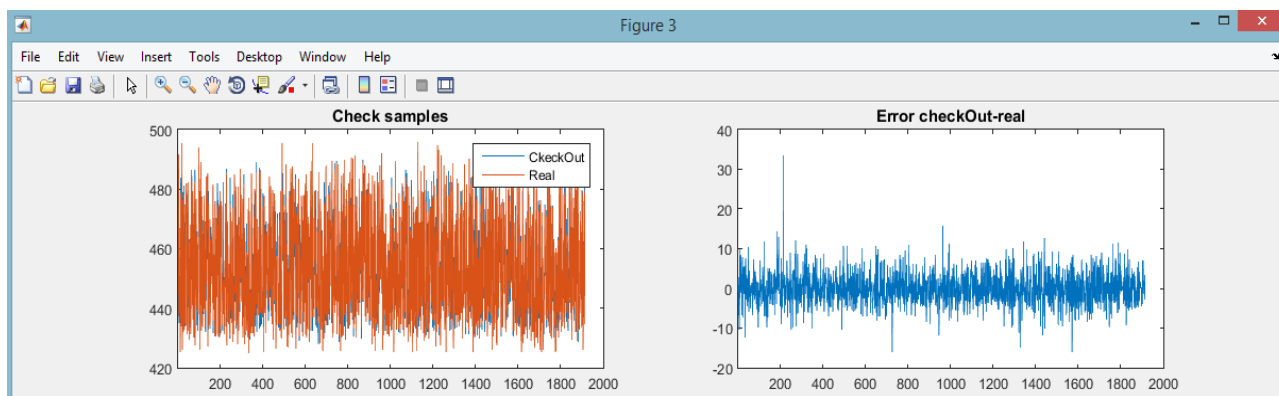


Check samples



Επιπλέον θα παρουσιάσουμε και το διάγραμμα σφάλματος για ολόκληρο το σύνολο check δεδομένων απλά για λόγους επίδειξης της σωστής συμπεριφοράς του μοντέλου πάνω σε ολόκληρο το σύνολο και όχι απλά σε ένα κομμάτι δεδομένων μεταξύ του 100 και 200. Έτσι έχουμε.

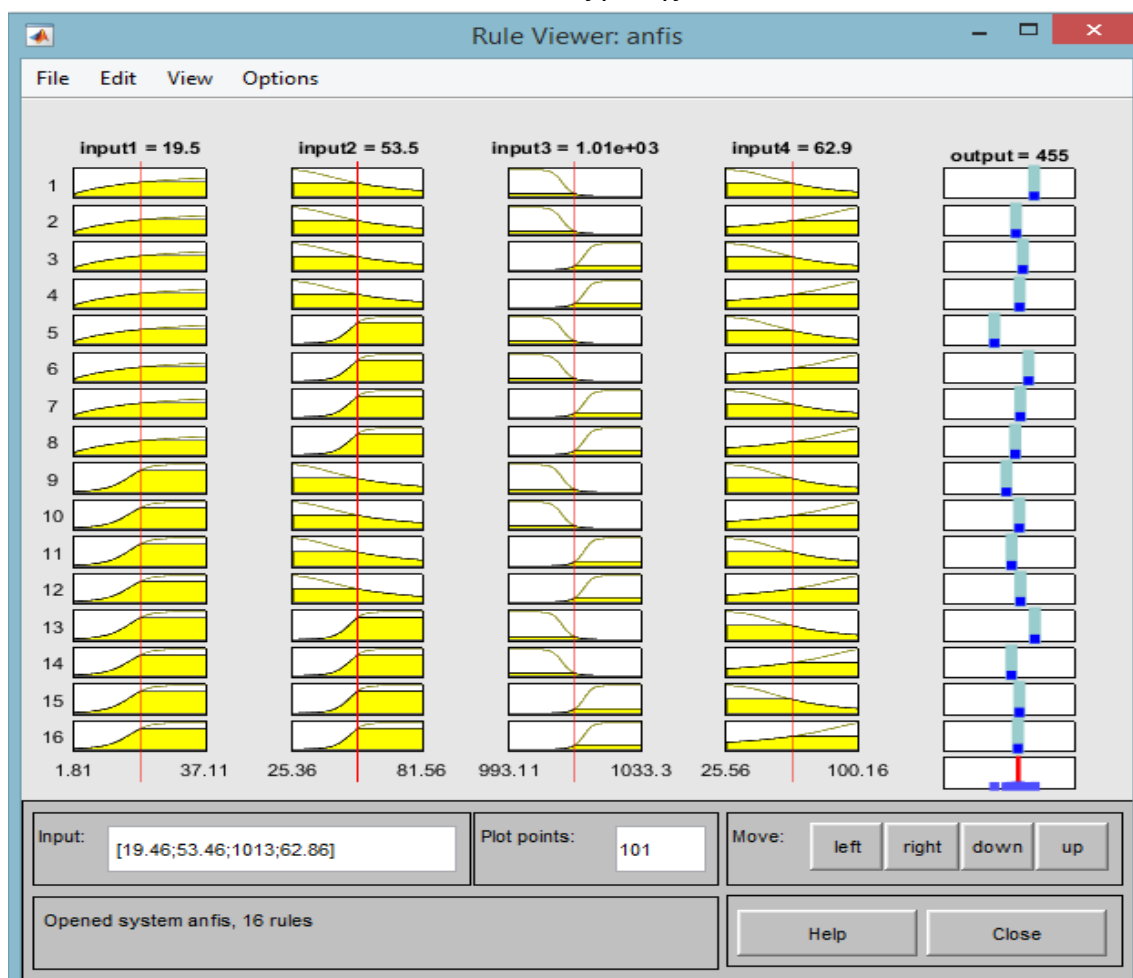
Ολόκληρο το σύνολο check δεδομένων



Όπως φαίνεται στο διάγραμμα το μοντέλο λειτουργεί για όλο το σύνολο δεδομένων.

Οι κανόνες που διαμορφώνονται για 4 εισόδους και 2 συναρτήσεις συμμετοχής είναι $2^4 = 16$ κανόνες. Για την παρουσίαση των κανόνων του μοντέλου καλούμε την fuzzy(tuned_fis) και επιλέγουμε από το GUI view rules.

Κανόνες βάσης



Πίνακας δεικτών απόδοσης

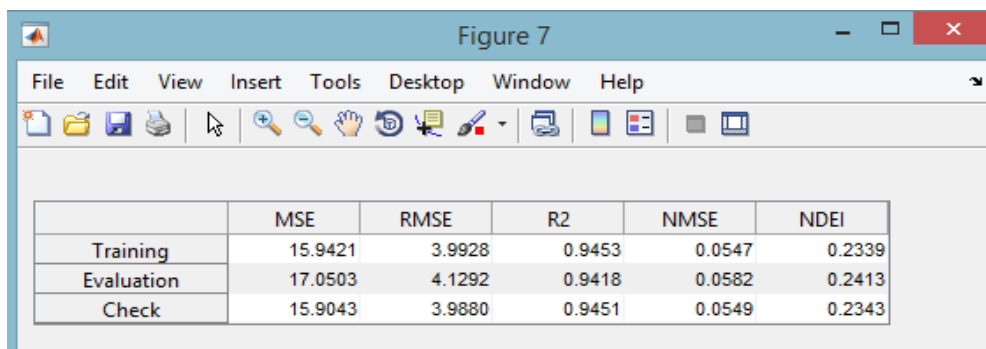


Figure 7 displays a table of performance metrics for the model. The table is titled 'Figure 7' and contains the following data:

	MSE	RMSE	R2	NMSE	NDEI
Training	15.9421	3.9928	0.9453	0.0547	0.2339
Evaluation	17.0503	4.1292	0.9418	0.0582	0.2413
Check	15.9043	3.9880	0.9451	0.0549	0.2343

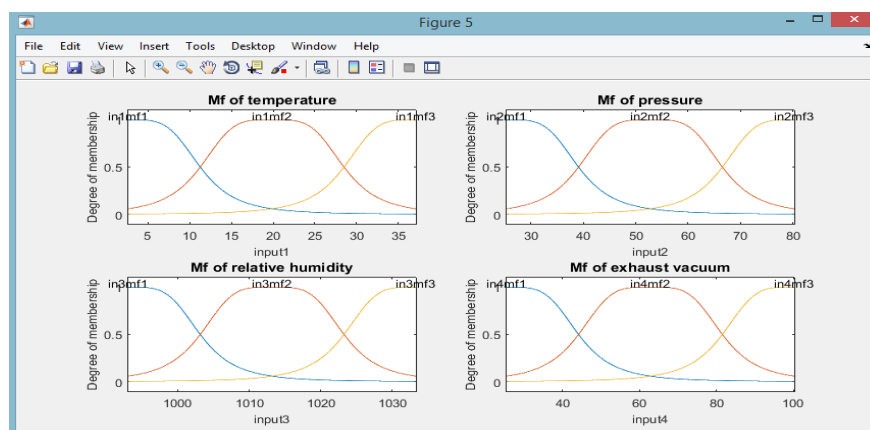
Σύμφωνα με τα αποτελέσματα ο μοντέλο με πολυωνυμική έξοδο λειτουργεί πολύ καλά με μέσο τετραγωνικό σφάλμα $MSE = 15.9$ ή θεωρώντας μία μέση απόλυτη τιμή 450 για τις τιμές τις εξόδου έχουμε ένα μέσο τετραγωνικό σφάλμα $15.9/450 = 0.03533$ ή 3.53% ,το οποίο είναι αποδεκτό.

Τέλος έχουμε και τα δεδομένα από το τέταρτο και τελευταίο μοντέλο.

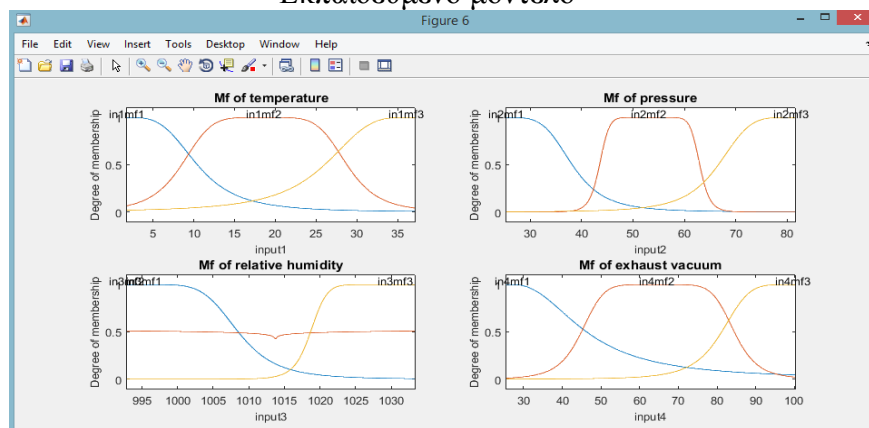
TSK model 4 με 3 συναρτήσεις συμμετοχής και έξοδο Polynomial

Η εκπαίδευση έχει γίνει για 300 εποχές. Τα δεδομένα πριν κάθε διαδικασία εκπαίδευσης ανακατεύονται για να μην έχουν την ίδια σειρά ώστε τα μοντέλα που προκύπτουν κάθε φορά να είναι περισσότερο αξιόπιστα.

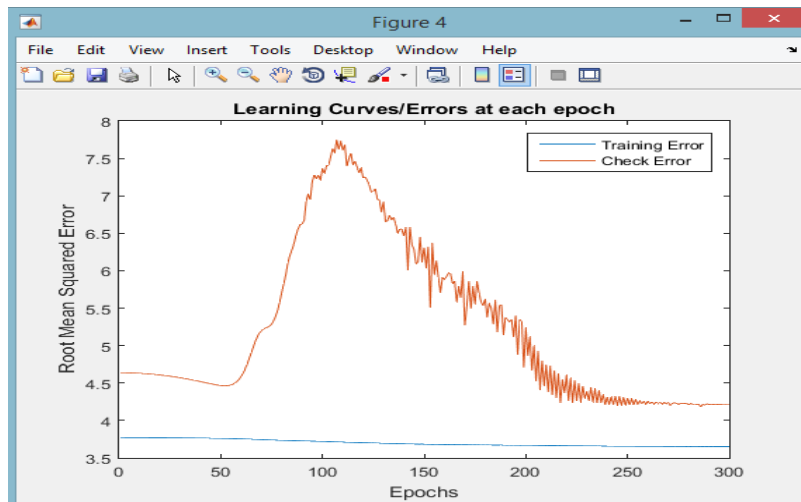
Αρχικό μοντέλο



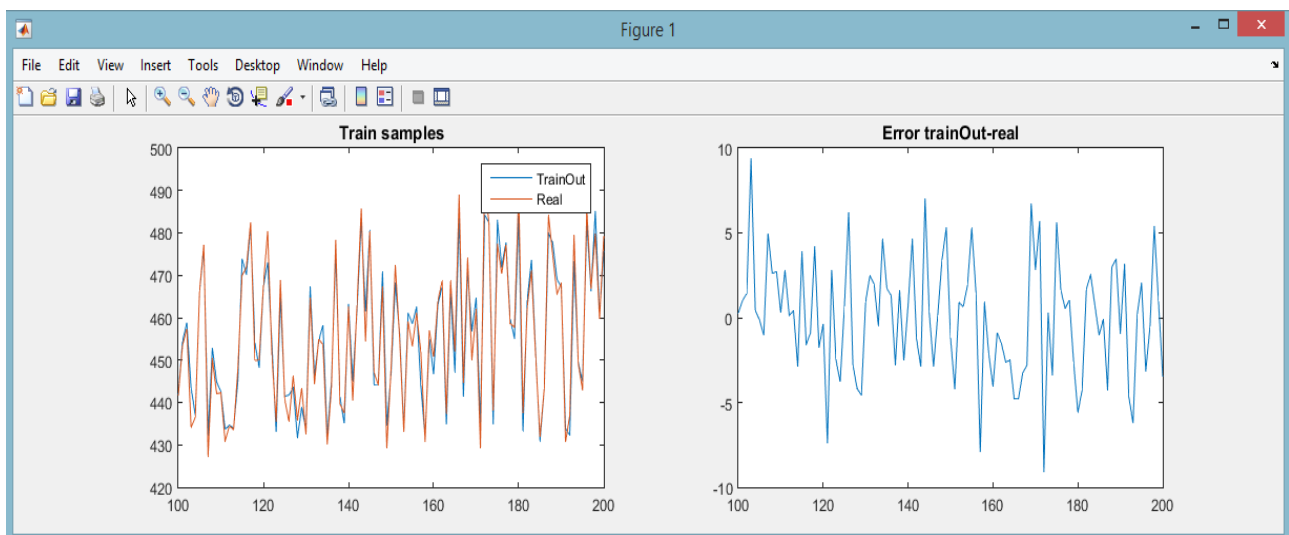
Εκπαιδευμένο μοντέλο



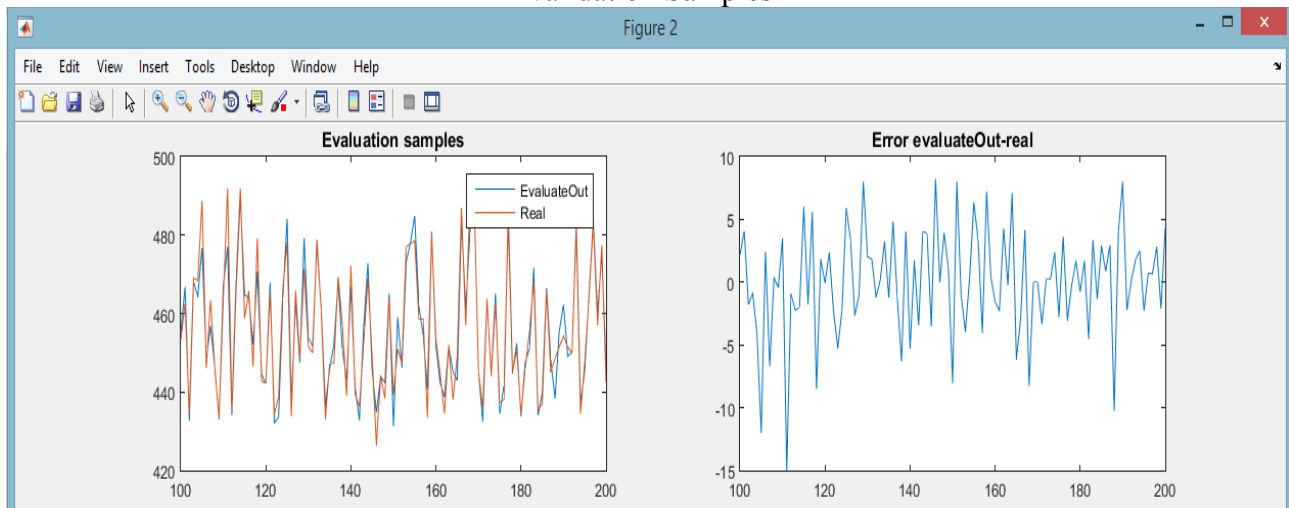
Διάγραμμα μάθησης (Learning Curves)



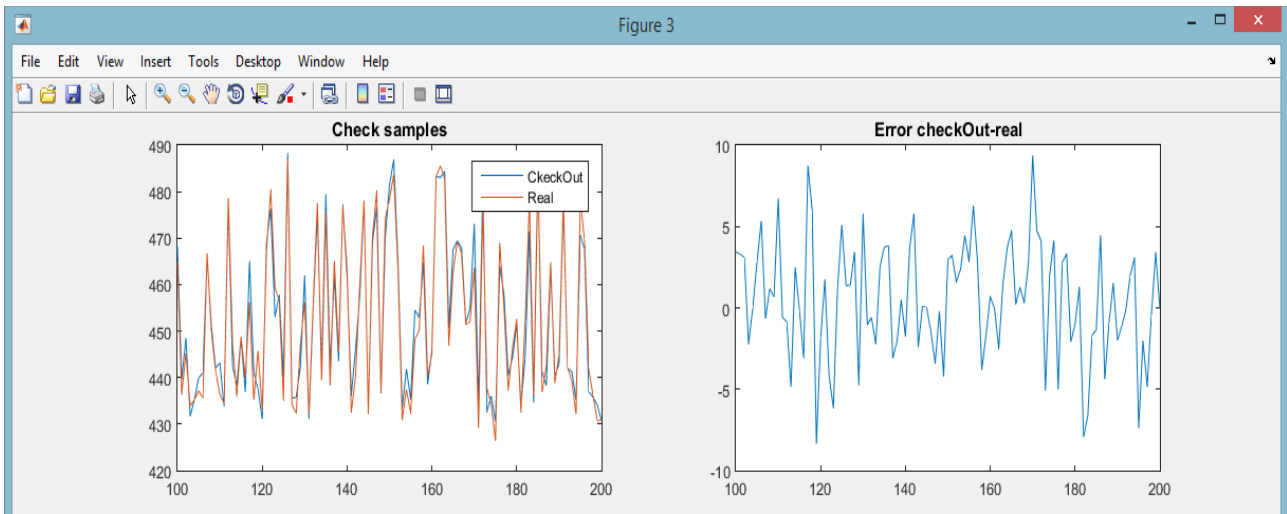
Training Sample



Validation Samples

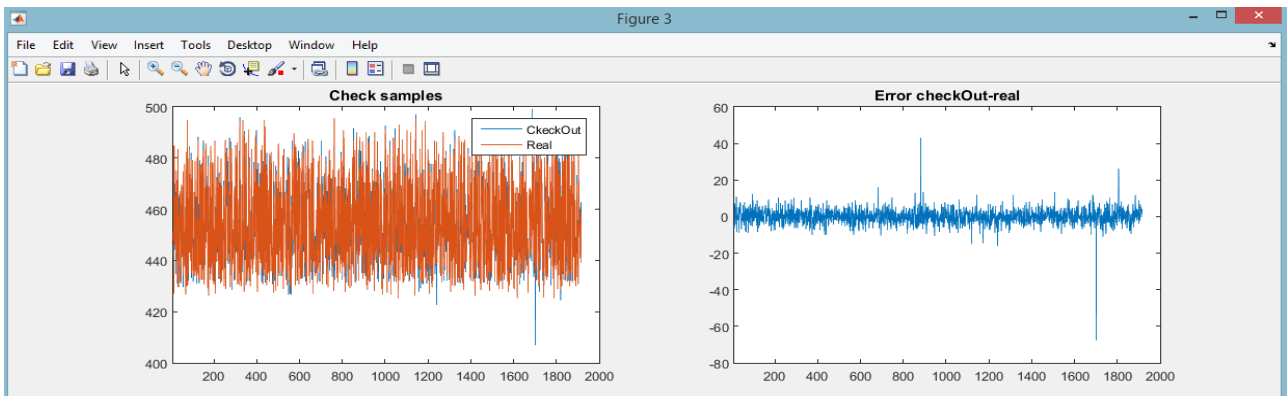


Check samples



Επιπλέον θα παρουσιάσουμε και το διάγραμμα σφάλματος για ολόκληρο το σύνολο check δεδομένων απλά για λόγους επίδειξης της σωστής συμπεριφοράς του μοντέλου πάνω σε ολόκληρο το σύνολο και όχι απλά σε ένα κομμάτι δεδομένων μεταξύ του 100 και 200. Έτσι έχουμε.

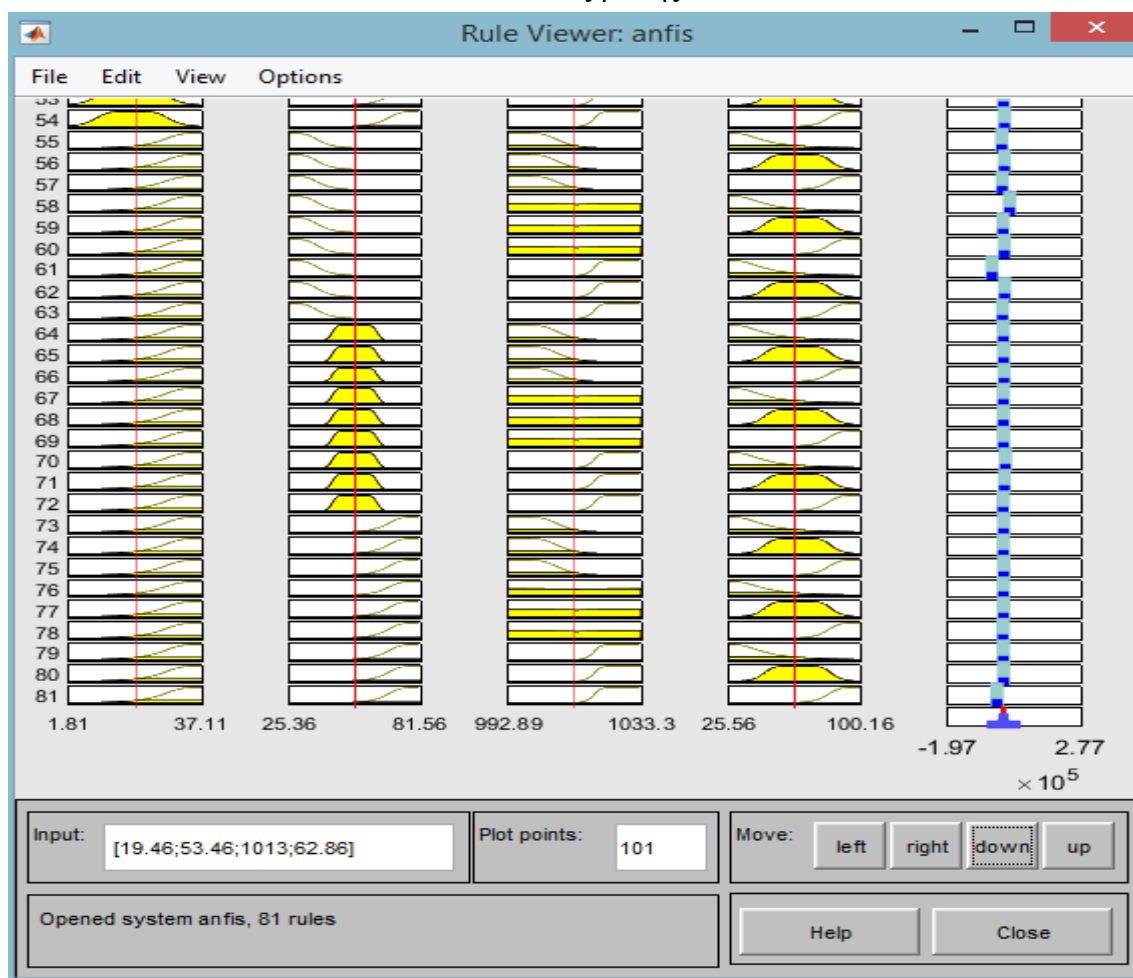
Ολόκληρο το σύνολο check δεδομένων



Όπως φαίνεται στο διάγραμμα το μοντέλο λειτουργεί για όλο το σύνολο δεδομένων.

Οι κανόνες που διαμορφώνονται για 4 εισόδους και 3 συναρτήσεις συμμετοχής είναι $3^4 = 81$ κανόνες. Για την παρουσίαση των κανόνων του μοντέλου καλούμε την fuzzy(tuned_fis) και επιλέγουμε από το GUI view rules.

Κανόνες βάσης



Πίνακας δεικτών απόδοσης

Figure 7

File Edit View Insert Tools Desktop Window Help

	MSE	RMSE	R2	NMSE	NDEI
Training	13.3286	3.6508	0.9547	0.0453	0.2129
Evaluation	17.8080	4.2200	0.9399	0.0601	0.2452
Check	17.3439	4.1646	0.9376	0.0624	0.2498

Συμπεράσματα και παρατηρήσεις

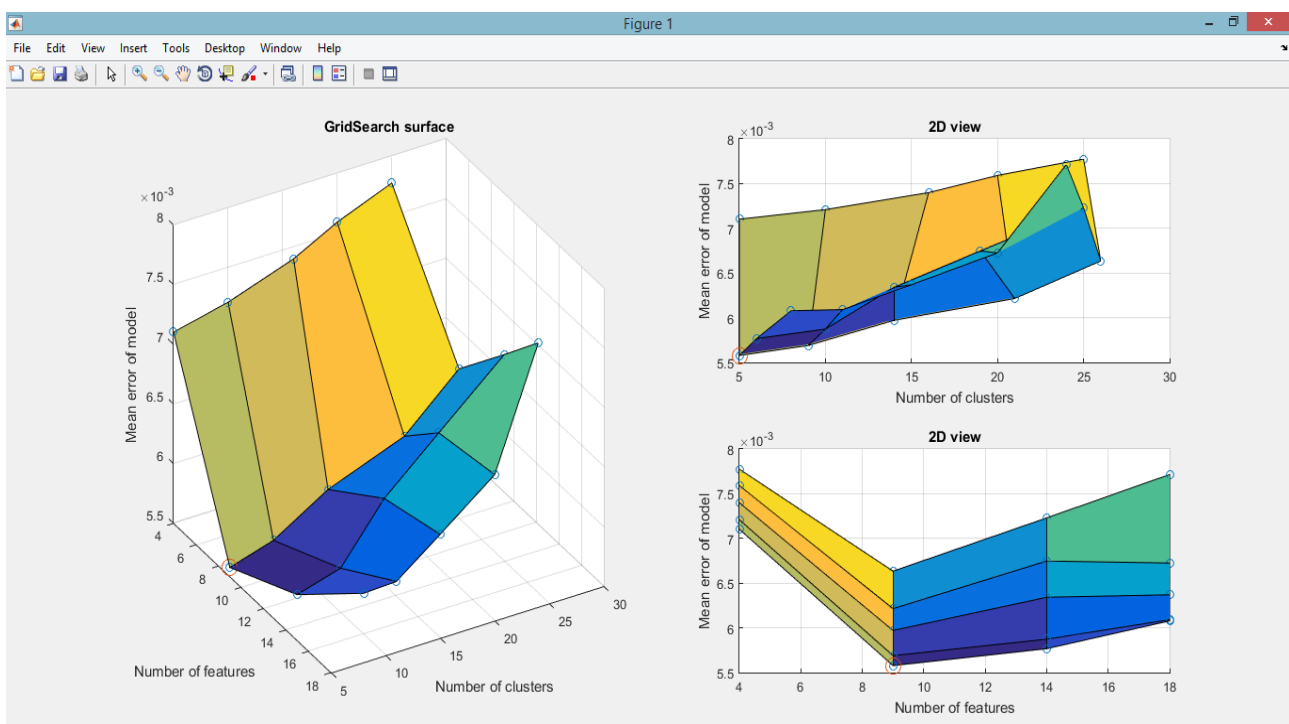
Όπως αναφέρθηκε και στην αρχή για την παρατήρηση της αποδοτικότητας των μοντέλων τα check δεδομένα είναι καταλληλότερα καθώς είναι άγνωστα για το μοντέλο και επομένως μπορούν να υποδείξουν αν το μοντέλο έχει εκπαιδευτεί σωστά και επίσης να υποδείξουν αν το μοντέλο έχει υποστεί υπερεκπαίδευση. Υπερεκπαίδευση συμβαίνει στην περίπτωση όπου το μοντέλο αντί να 'μάθει' να κάνει προβλέψεις για την έξοδο με βάση τις τιμές της εισόδου, εκπαιδεύεται σε τέτοιο βαθμό που προσαρμόζεται στα δεδομένα εκπαίδευσης (training) και χάνει την ικανότητα του να κάνει προβλέψεις και για άλλα δεδομένα, κάτι το οποίο φυσικά δεν είναι επιθυμητό. Έτσι με χρήση των check δεδομένων μπορούμε να σιγουρέψουμε την γενικότητα του μοντέλου. Επιπλέον για τον ίδιο σκοπό χρησιμοποιούνται κατά την διάρκεια της εκπαίδευσης τα evaluation δεδομένα τα οποία δείχνουν αν το μοντέλο έχει υποστεί υπερεκπαίδευση συναρτήσει των εποχών. Από τα διαγράμματα μάθησης (Learning Curves) των παραπάνω μοντέλων μπορούμε εύκολα να συμπεράνουμε αν τα μοντέλα έχουν υποστεί υπερεκπαίδευση κατά την διάρκεια εκπαίδευσης παρατηρώντας τις πορτοκαλί γραφικές παραστάσεις που αντιστοιχούν στο μέσο τετραγωνικό σφάλμα των evaluation/validation δεδομένων. Για το μοντέλο 1, 2 και 3 βλέπουμε ότι δεν εμφανίζεται υπερεκπαίδευση αφού στο τέλος της εκπαίδευσης έχουμε καταλήξει σε ελάχιστο σφάλμα για τα δεδομένα επικύρωσης. Το μοντέλο 2 εμφανίζει ένα τοπικό ελάχιστο στην περιοχή των 100 εποχών, δηλαδή μετά από αυτό το σημείο το σφάλμα αυξάνεται και το οποίο δείχνει ότι το μοντέλο αρχίζει να υπερεκπαδεύεται, παρόλα αυτά στην συνέχεια της εκπαίδευσης το μοντέλο συνεχίζει την εκπαίδευση του και το σφάλμα μειώνεται φτάνοντας τελικά στην τελική ελάχιστη τιμή του στις 300 εποχές. Για το μοντέλο 4 παρατηρούμε μία ανεπιθύμητη συμπεριφορά, όπως φαίνεται από το διάγραμμα μάθησης καθώς ξεκινάει η διαδικασία εκπαίδευσης αρχικά το σφάλμα μειώνεται αλλά περίπου μετά τις 50 εποχές το μοντέλο ξεκινάει να υπερεκπαιδεύεται, αφού όπως μπορούμε να παρατηρήσουμε το σφάλμα από τα training data (μπλε γραφική) συνεχίζει να μειώνεται από εποχή σε εποχή από την άλλη μεριά όμως το σφάλμα από τα δεδομένα επικύρωσης αυξάνεται. Αυτό υποδεικνύει ότι το μοντέλο χάνει την ικανότητα του να κάνει προβλέψεις για άλλα δεδομένα εκτός των training δεδομένων. Τελικά μετά από τις 100 εποχές και μέχρι τις 250 το μοντέλο μειώνει το σφάλμα πρόβλεψης κάτω από την αρχική του τιμή και το πρόβλημα της υπερεκπαίδευσης εξαφανίζεται. Το τελικό σφάλμα όμως δεν είναι αρκετά μικρό, το οποίο τελικά δείχνει ότι η αύξηση του αριθμού των συναρτήσεων συμμετοχής μπορεί όντως να δημιουργήσει πρόβλημα υπερεκπαίδευσης. Σχετικά με την μορφή των εισόδων βλέπουμε ότι τα μοντέλα προσαρμόζουν τις συναρτήσεις συμμετοχής πολύ περισσότερο σε κάποιες από τις εισόδους σε σχέση με τις υπόλοιπες σε τέτοιο βαθμό ώστε κάποιες από τις συναρτήσεις συμμετοχής να καταλαμβάνουν πολύ μεγάλο μέρος του χώρου εισόδου, ενώ κάποιες άλλες να περιορίζονται σε κάποια περιοχή του χώρου. Επιπλέον κάποιες από τις συναρτήσεις εμφανίζουν πολύ μεγαλύτερο βαθμό συμμετοχή από κάποιες άλλες. Η περιοχή που καταλαμβάνεται καθώς και ο βαθμός συμμετοχής προφανώς καθορίζεται από την υβριδική μέθοδο εκμάθησης του μοντέλου και από το πόσο σημαντική είναι η είσοδος για την σωστή πρόβλεψη της εξόδου. Γενικά βλέπουμε ότι όλα τα μοντέλα προσεγγίζουν την έξοδο σε ικανοποιητικό βαθμό, παρόλα αυτά από τους πίνακες των δεικτών απόδοσης φαίνεται ότι τα μοντέλα με πολυωνυμική έξοδο κάνουν καλύτερη πρόβλεψη σε σχέση με αυτά των σταθερών εξόδων, αφού έχουν περισσότερες παραμέτρους (ως γραμμικές πολυμεταβλητές συναρτήσεις) και εισάγουν μεγαλύτερη πολυπλοκότητα στο σύστημα, άρα και καλύτερη πρόβλεψη. Ο αριθμός των συναρτήσεων συμμετοχής όσο αυξάνεται μειώνει το σφάλμα πρόβλεψης αφού όπως φαίνεται το μοντέλο 2 έκανε καλύτερη πρόβλεψη από το μοντέλο 1, παρόλα αυτά η αύξηση του αριθμού των συναρτήσεων συμμετοχής εισάγει το πρόβλημα της υπερεκπαίδευσης και αυτό φαίνεται στο μοντέλο 4. Τελικά το μοντέλο με τα καλύτερα αποτελέσματα είναι το μοντέλο 3 με μέσο τετραγωνικό σφάλμα $MSE = 15.9$ και $R^2 = 0.9451$.

2)Εφαρμογή σε dataset με υψηλή δραστηριότητα

Στην συνέχεια της εργασίας αντί για την κλασική περίπτωση του grid partitioning του χώρου εισόδου μια εναλλακτική επιλογή είναι η χρήση μεθόδων ομαδοποίησης για το διαχωρισμό του χώρου εισόδου και η αρχικοποίηση των ασαφών συνόλων πάνω στις ομάδες που προέκυψαν. Για αυτό τον σκοπό θα χρησιμοποιήσουμε την μέθοδο της αφαιρετικής ομαδοποίησης Subtractive Clustering (SC – genfis2). Αρχικά θα διαχωρίσουμε τα δεδομένα σε τρεις ομάδες δεδομένων όπως και πριν, δηλαδή σε training, evaluation και check data. Στην συνέχεια για να επιλέξουμε τα καλύτερα χαρακτηριστικά του μοντέλου με χρήση της μεθόδου της 5-πτυχης διασταυρωμένης επικύρωσης (5-fold cross validation) θα χρησιμοποιήσουμε το σύνολο των training δεδομένων αφού πρώτα τα χωρίσουμε σε ένα ποσοστό 80% και 20% για την διαδικασία της εκπαίδευσης και επικύρωσης. Πριν εκτελεστεί η διαδικασία της διασταυρωμένης επικύρωσης θα καλέσουμε την συνάρτηση relief με είσοδο τα training δεδομένα ώστε να εκτελεστεί ο αλγόριθμος relief και να βρεθούν οι καλύτερες εισοδοί/predictors. Στην συνέχεια ανάλογα με τον αριθμό των χαρακτηριστικών NF που επιλέγονται για την διαδικασία του grid search θα επιλέγονται και οι πρώτες NF εισοδοί που προέκυψαν από τον αλγόριθμο relief. Οι ελεύθερες μεταβλητές που επιλέχθηκαν για το grid search είναι: α)για τον αριθμό των χαρακτηριστικών NF = [4 9 14 18] ενώ β)για την περίπτωση των cluster/κανόνων δεν μπορούμε να έχουμε ακριβείς αριθμό, αλλά μέσα από τον κατάλληλο καθορισμό των ακτίνων των cluster φροντίζουμε να πάρουμε τιμές σε διάστημα μεταξύ του 5 και 25 με τιμές κοντά στους αριθμούς NR = [5,10,15,20,25]. Μεγαλύτερες τιμές για τον αριθμό των χαρακτηριστικών και των αριθμών των κανόνων δεν έχουν επιλεγεί λόγω του μεγάλου χρόνου εκτέλεσης που απαιτεί η διαδικασία εκπαίδευσης.

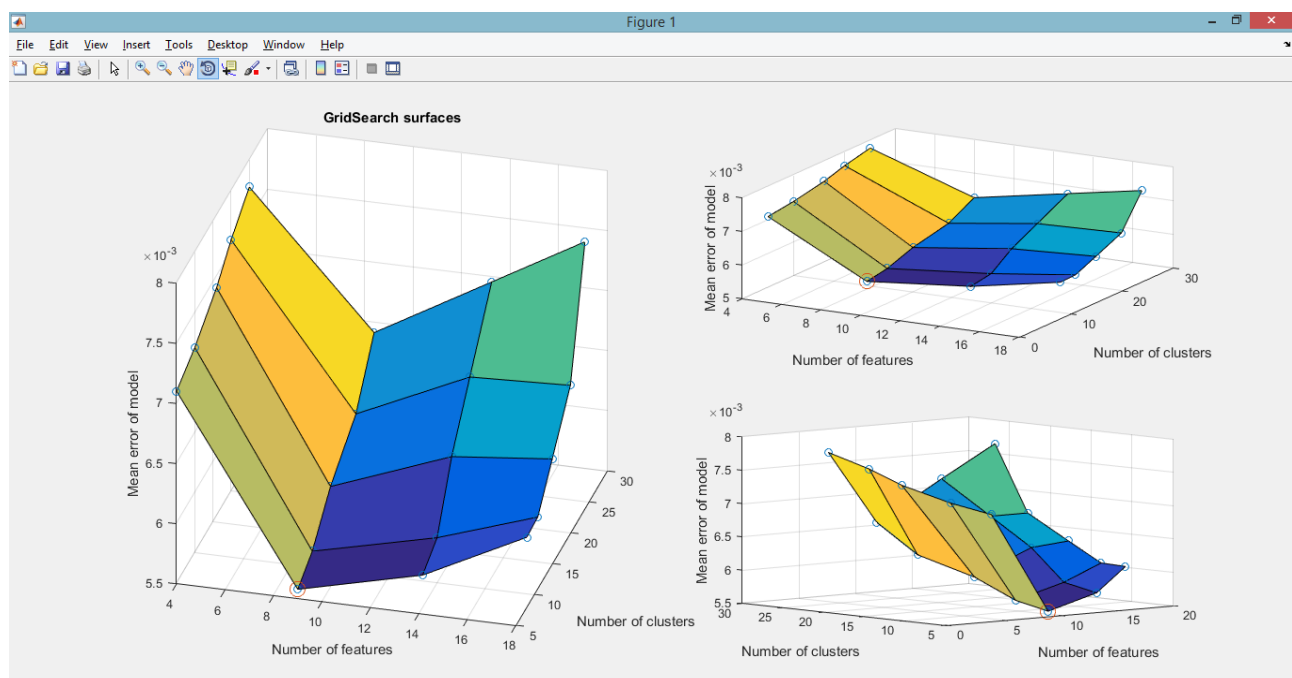
Αναζήτηση πλέγματος (Grid Search)

Η διάσταση του πλέγματος είναι 4x5 δηλαδή επιλέγουμε 4 διαφορετικούς αριθμούς χαρακτηριστικών και 5 διαφορετικές ακτίνες για κάθε αριθμό χαρακτηριστικών και η εκπαίδευση έχει γίνει για 100 εποχές. Εκτελώνοντας επομένως την 5-fold διασταυρωμένη επικύρωση για κάθε συνδυασμό των ελεύθερων μεταβλητών προκύπτει το παρακάτω διάγραμμα στο οποίο φαίνεται το μέσο σφάλμα της 5-fold διασταυρωμένης επικύρωσης συναρτήσει του αριθμού των χαρακτηριστικών και του αριθμού των κανόνων σε μορφή 3D απεικόνισης (αριστερά).



Στο παραπάνω διάγραμμα στα δεξιά φαίνονται οι πλάγιες προβολές της επιφάνειας. Τα σημεία με κύκλους αντιστοιχούν στις πραγματικές τιμές των δεδομένων NF και NR που αντιστοιχούν στο πλέγμα ενώ μεταξύ αυτών των στοιχείων σχεδιάζεται μία επιφάνεια δίνοντας με αυτό τον τρόπο μια προσέγγιση της πραγματικής επιφάνειας του πλέγματος που προκύπτει για κάθε αριθμό χαρακτηριστικών και κανόνων. Η επιφάνεια δείχνει την συμπεριφορά του μοντέλου με βάση τις ελεύθερες μεταβλητές. Το σημείο το οποίο είναι μέσα στον κόκκινο κύκλο αντιστοιχεί στο συνδυασμό των ελεύθερων μεταβλητών με το ελάχιστο μέσο σφάλμα, δηλαδή στο καλύτερο μοντέλο.

Παρακάτω φαίνεται ένα επιπλέον διαγράμματα του Grid Search από διαφορετικές οπτικές γωνίες για καλύτερη παρατήρηση.



Σχόλια

Από τα παραπάνω διαγράμματα του πλέγματος βλέπουμε ότι το μέσο σφάλμα μειώνεται καθώς αυξάνεται ο αριθμός των χαρακτηριστικών μέχρι τα 9 χαρακτηριστικά περίπου και από εκεί και πέρα αυξάνεται. Αυτό μπορεί να συμβαίνει λόγω του ότι τα 9 χαρακτηριστικά πιθανόν περιγράφουν την έξοδο με τον καλύτερο δυνατό τρόπο, ενώ τα παραπάνω χαρακτηριστικά εισάγουν επιπλέον πολυπλοκότητα που το μοντέλο αδυνατεί να αποκωδικοποιήσει με αποτέλεσμα να εισάγεται επιπλέον σφάλμα στην έξοδο. Να σημειωθεί ότι σε ένα μοντέλο νευρωνικού δικτύου οι περισσότεροι είσοδοι δεν συνεπάγονται και καλύτερο αποτέλεσμα πρόβλεψης, γενικά τα χαρακτηριστικά ενός νευρωνικού εξαρτώνται από τα δεδομένα και επομένως ο καθορισμός των χαρακτηριστικών συνδέεται στενά με την διαδικασία εκπαίδευσης του δικτύου. Επιπλέον το σφάλμα συναρτήσει του αριθμού των κανόνων αυξάνεται όσο αυξάνεται και ο αριθμός των κανόνων, αυτό πιθανόν συμβαίνει λόγω υπερεκπαίδευσης του μοντέλου και προσαρμογής του πάνω στα δεδομένα εκπαίδευσης. Οι περιοχές με το μεγαλύτερο σφάλμα αντιστοιχούν στην περιοχή με πράσινο χρώμα. Οι έντονες μπλε περιοχές του πλέγματος αντιστοιχούν στις μικρότερες τιμές σφάλματος, μέσα στις οποίες βρίσκονται προφανώς και οι βέλτιστες μεταβλητές του μοντέλου μας. Σύμφωνα με το παραπάνω διάγραμμα πλέγματος προκύπτει ότι:

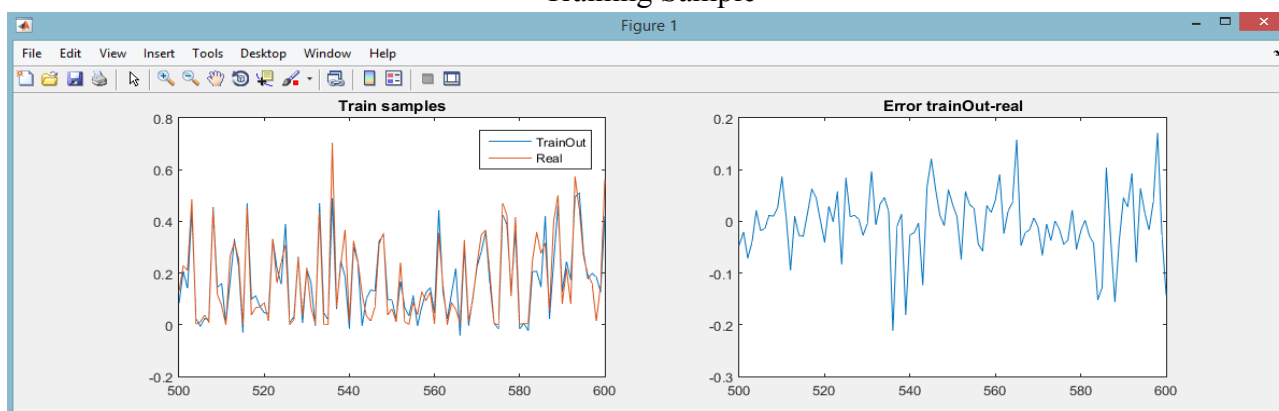
Οι βέλτιστες τιμές του μοντέλου είναι η χρήση NF = 9 χαρακτηριστικών και NR= 5 κανόνων (Cluster radius: 0.830).

Εφόσον έχουμε βρει τις βέλτιστες τιμές παραμέτρων συνεχίζουμε την ανάλυση δημιουργώντας ένα μοντέλο με τις ίδιες παραμέτρους και το εκπαιδεύουμε με χρήση των training data και evaluation data για τον έλεγχο υπερεκπαίδευσης. Μετά την εκπαίδευση χρησιμοποιούμε το check δεδομένα για την μελέτη της απόδοσης του μοντέλου όπως ακριβώς κάναμε και στο πρώτο κομμάτι της εργασίας. Τα αποτελέσματα από την εκπαίδευση του μοντέλου είναι:

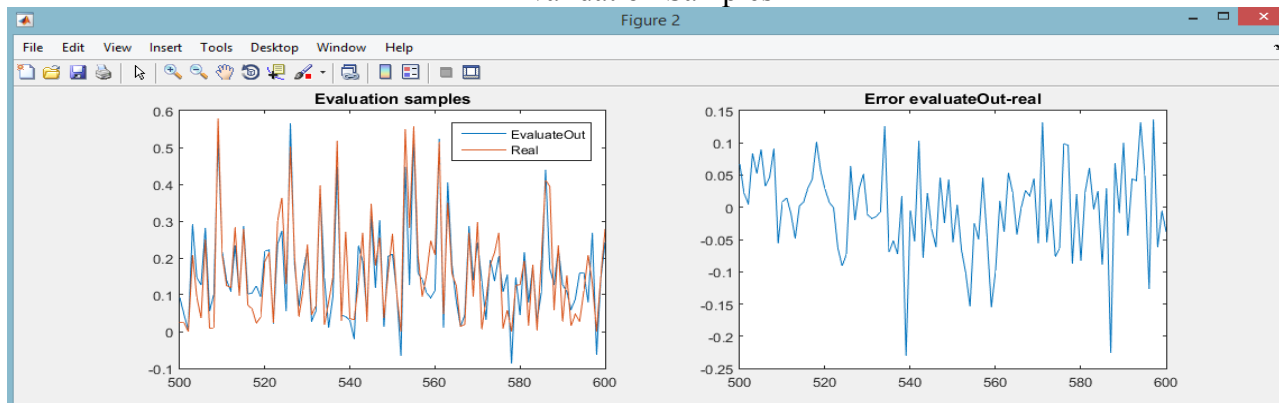
Βέλτιστο μοντέλο

Η εκπαίδευση του μοντέλου έχει γίνει για 600 εποχές για την παρατήρηση τυχόν υπερεκπαίδευσης του μοντέλου, τα διαγράμματα που απεικονίζουν το σφάλμα και την πραγματική τιμή της εξόδου για ένα διάστημα 100 τιμών για τις τρεις κατηγορίες δεδομένων είναι:

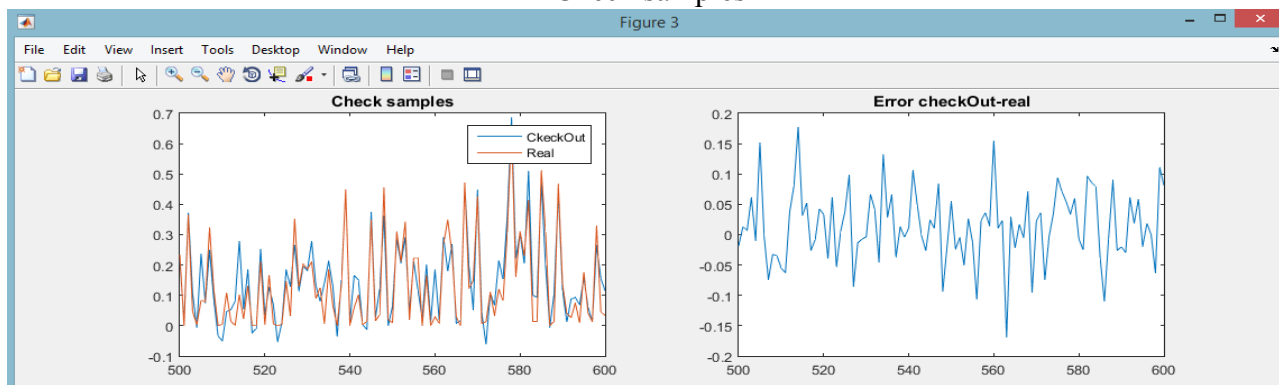
Training Sample



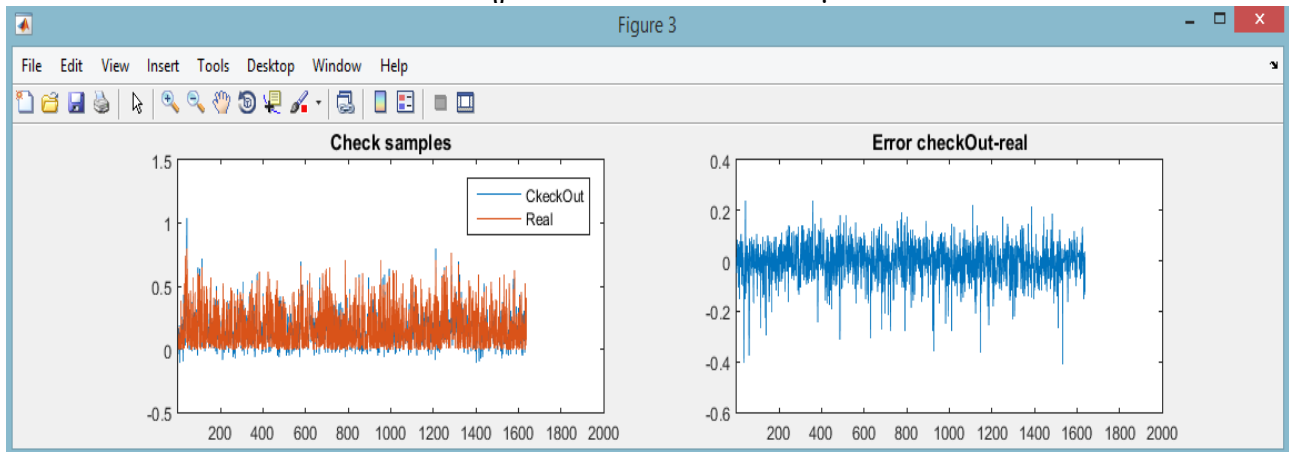
Validation Samples



Check samples



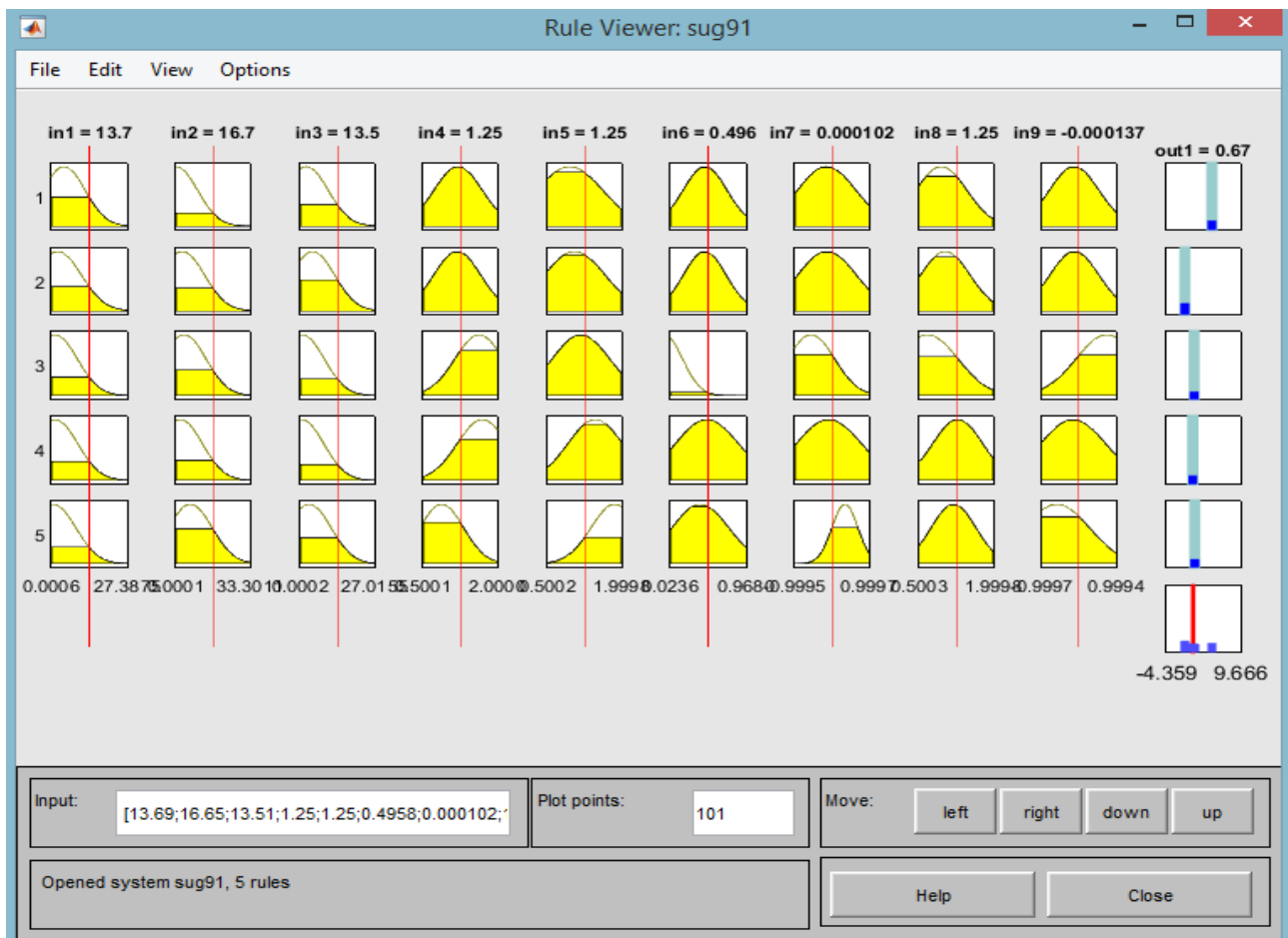
Ολόκληρο το σύνολο check δεδομένων



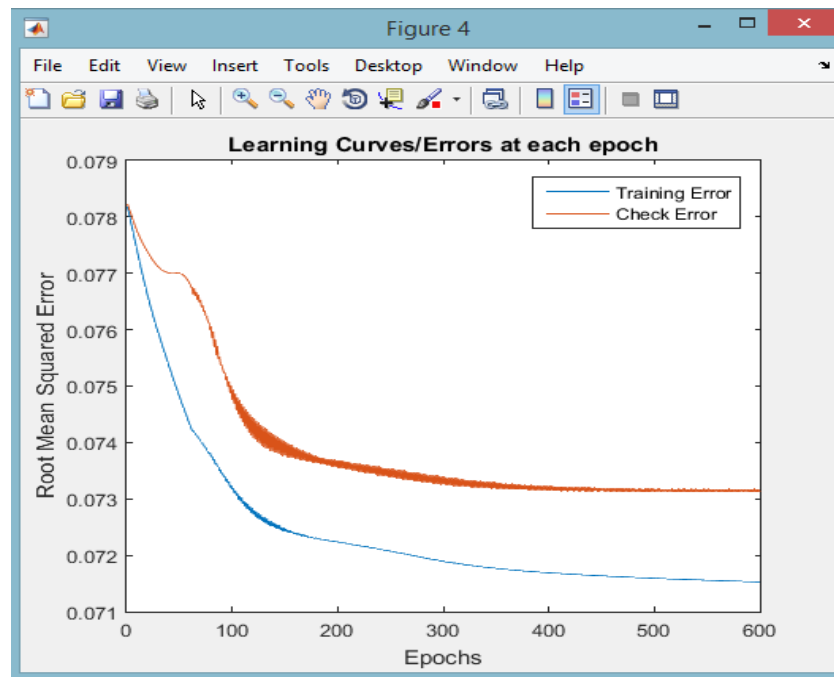
Τα διαγράμματα σφάλματος για τα training και validation δεδομένα παρουσιάζονται για λόγους πληρότητας ώστε να φανεί η συμπεριφορά του μοντέλου πάνω σε ολόκληρο το δείγμα δεδομένων. Μεγαλύτερης σημασίας είναι τα αποτελέσματα που προκύπτουν από τα check δεδομένα καθώς αυτά είναι δεδομένα άγνωστα για το μοντέλο.

Στην συνέχεια παρουσιάζονται οι 5 κανόνες που διαμορφώνονται για 9 εισόδους. Για την παρουσίαση των κανόνων του μοντέλου καλούμε την `fuzzy(tuned_fis)` και επιλέγουμε από το GUI `view rules`, έτσι παίρνουμε.

Κανόνες βάσης

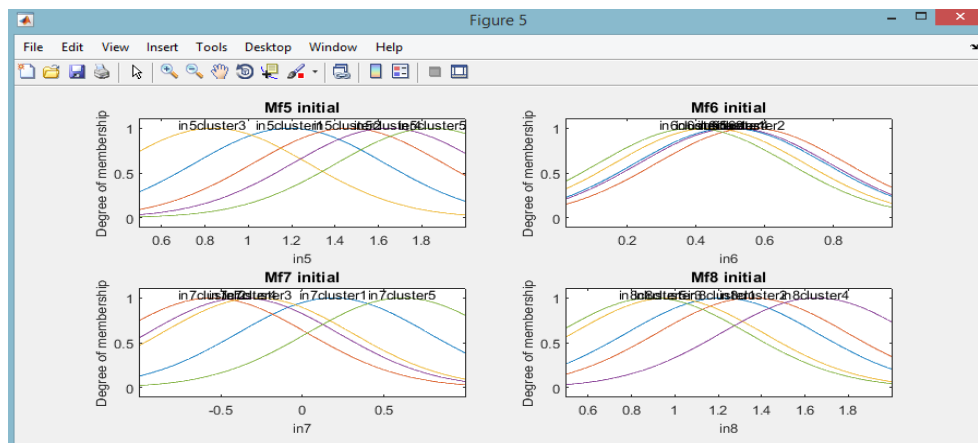


Διάγραμμα μάθησης (Learning Curves)

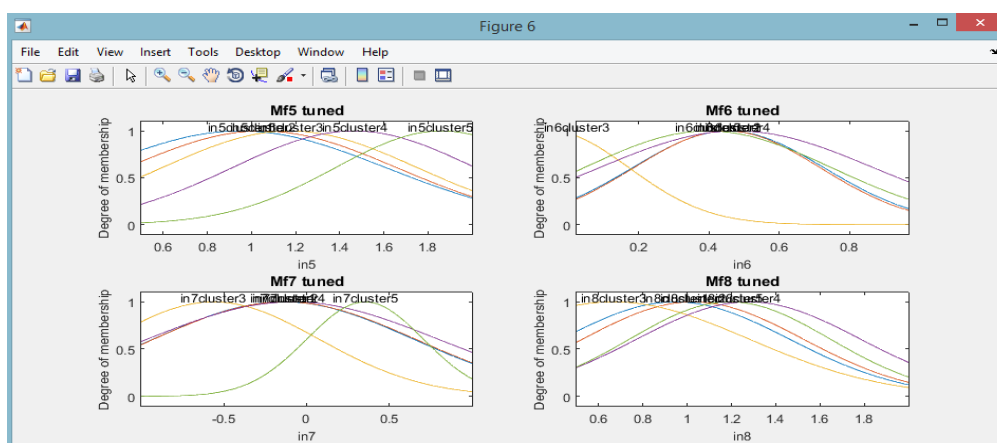


Τα ασαφή σύνολα για τις εισόδους 5,6,7,8 πριν και μετά την εκπαίδευση φαίνονται στα παρακάτω διαγράμματα.

Αρχικό μοντέλο

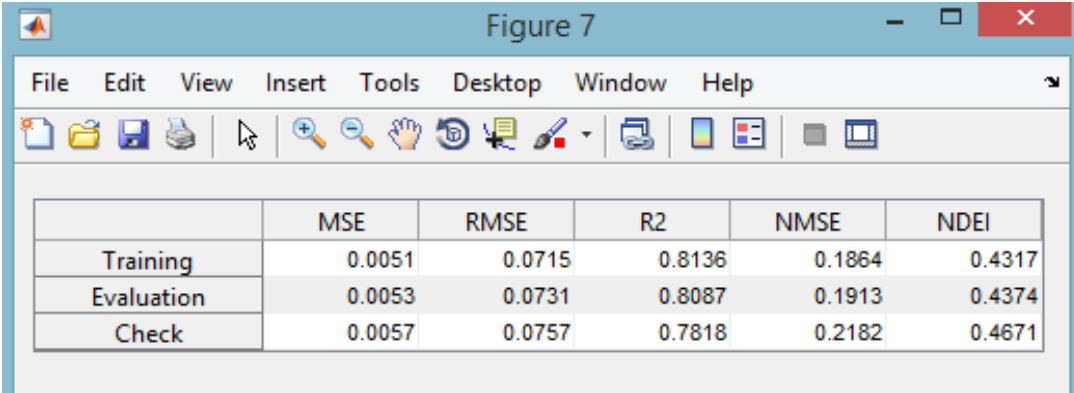


Εκπαιδευμένο μοντέλο



Τέλος παρουσιάζουμε τους δείκτες απόδοσης που αναφέρονται στην εκφώνηση. Με χρήση της συνάρτησης `unitable` του Matlab έχουμε τοποθετήσει τους δείκτες σε πίνακα για την καλύτερη παρουσίαση.

Πίνακας δεικτών απόδοσης



The screenshot shows a MATLAB window titled 'Figure 7' with a standard menu bar (File, Edit, View, Insert, Tools, Desktop, Window, Help) and a toolbar. Below the toolbar is a table with performance metrics for Training, Evaluation, and Check datasets.

	MSE	RMSE	R2	NMSE	NDEI
Training	0.0051	0.0715	0.8136	0.1864	0.4317
Evaluation	0.0053	0.0731	0.8087	0.1913	0.4374
Check	0.0057	0.0757	0.7818	0.2182	0.4671

Συμπεράσματα

Από τα διαγράμματα σφάλματος βλέπουμε ότι το μοντέλο με τις βέλτιστες παραμέτρους τελικά μπορεί να κάνει σωστές προβλέψεις για όλες τις ομάδες δεδομένων (training, validation, check). Από το πίνακα δεικτών απόδοσης βλέπουμε ένα $MSE = 0.0057$ για τα check δεδομένα. Συγκριτικά με τις απόλυτες τιμές της εξόδου που όπως φαίνεται από τα διαγράμματα σφάλματος κυμαίνονται περίπου γύρω από 0.4, το παραπάνω μέσω τετραγωνικό σφάλμα είναι περίπου ένα σφάλμα $0.0057/0.4 = 0.01425$ ή 1.425% συγκριτικά με την μέση απόλυτη τιμή των δεδομένων εξόδου. Επιπλέον από τους συντελεστές R^2 και NMSE και NDEI βλέπουμε ότι η εκτίμηση είναι πολύ κοντά στην πραγματική τιμή αφού όσο μεγαλύτερος και κοντά στην μονάδα είναι ο συντελεστής προσδιορισμού R^2 και μικρότεροι οι συντελεστές NMSE και NDEI τόσο πιο κοντά είναι η εκτίμηση του μοντέλου στην πραγματική τιμή της εξόδου. Βέβαια η εκτίμηση δεν είναι απόλυτα σωστή αλλά όπως φαίνεται και από τα διαγράμματα ακολουθεί πολύ καλά τις μεταβολές της εξόδου (ανεβοκατεβάσματα). Συμπερασματικά, δεδομένου της χαμηλής πολυπλοκότητας του μοντέλου, αφού χρησιμοποιούνται μόνο οι 9 είσοδοι από τις 32 και σχηματίζονται μόνο 5 κανόνες, μπορούμε να συμπεράνουμε ότι η εκτίμηση του μοντέλου είναι πολύ καλή. Σε αντίθετη περίπτωση που θα έπρεπε να χρησιμοποιήσουν grid partitioning με δύο ή τρία ασαφή σύνολα ανά είσοδο τότε ο αριθμός των κανόνων θα ήταν 2^9 για τα 2 ασαφή σύνολα ανά είσοδο και 3^9 για 3 ασαφή σύνολα ανά είσοδο για τον ίδιο αριθμό χαρακτηριστικών, αν επιπλέον δεν κάναμε ούτε επιλογή χαρακτηριστικών και χρησιμοποιούσαμε και τα 32 χαρακτηριστικά τότε θα είχαμε 2^{32} για τα 2 ασαφή σύνολα ανά είσοδο και 3^{32} για 3 ασαφή σύνολα ανά είσοδο. Ο αριθμός των κανόνων είναι απαγορευτικά μεγάλος καθώς η πολυπλοκότητα και ο χρόνος εκτέλεσης της εκπαίδευσης θα ήταν πολύ μεγάλος σε σημείο να μην μπορεί να εφαρμοστεί πρακτικά μέσω ενός απλού υπολογιστικού συστήματος. Αρά η επιλογή των χαρακτηριστικών και η εφαρμογή του grid search αποτελεί πολύ καλή λύση για την επίλυση του προβλήματος και την κατασκευή ενός μοντέλου πρόβλεψης για πρακτική εφαρμογή.

Περιγραφή αρχείων εργασίας

- **TSK_hybr_models.m** : Το αρχείο περιέχει το κώδικα για το πρώτο κομμάτι της εργασίας για τα 4 μοντέλα TSK.
- **reliefCall.m** : Το script καλείται για να εκτελέσει την συνάρτηση relief ώστε γίνει η επιλογή των κατάλληλων χαρακτηριστικών/predictors. Κατά την εκτέλεση του script αποθηκεύετε η σειρά των καλύτερων χαρακτηριστικών αφού αυτή δεν αλλάζει, έτσι ώστε να μην εκτελείται συνέχεια ο αλγόριθμος σε επαναλαμβανόμενες δοκιμές στα πλαίσια της εργασίας.
- **findClustersNum.m** : Η συνάρτηση χρησιμοποιείται για τον υπολογισμό των αριθμών των cluster συναρτήσει του αριθμού των χαρακτηριστικών και της ακτίνας των cluster. Η συνάρτηση είναι βοηθητική και δεν βρίσκεται στα ζητούμενα της εργασίας απλά χρησιμοποιείται κατά την διαδικασία του καθορισμού των ακτίνων των cluster. Αν για παράδειγμα την καλέσουμε με όρισμα των αριθμό των χαρακτηριστικών και την ακτίνα των cluster θα εκτυπώσει τον αριθμό των cluster/κανόνων του μοντέλου, δηλαδή αν καλέσουμε findClustersNum(4, [0.12 0.23 03]) θα εκτυπώσει 3 αριθμούς που αντιστοιχούν στους αντίστοιχους αριθμούς των clusters. Έτσι υπολογίζουμε ακτίνες ώστε να μας δίνουν αριθμό κανόνων κοντά στις τιμές [5 10 15 20 25] όπως ζητείται στην εκφώνηση.
- **gridSearch.m** : Εκτελεί τον κώδικα που αντιστοιχεί για την grid search 5-fold cross validation διαδικασία.
- **plotGridSearch.m** : Καλείται για να δημιουργήσει την γραφική παράσταση της επιφάνειας του πλέγματος.
- **bestTSKmodel** : Καλείται για την εκπαίδευση του καλύτερου μοντέλου. Αφού εισάγουμε τις παραμέτρους για το καλύτερο μοντέλο στην συνέχεια τρέχουμε το script για να εκπαιδευτεί το μοντέλο.
- **Αρχεία δεδομένων** : Τα διάφορα αρχεία δεδομένων που περιέχονται μέσα στο φάκελο περιέχουν τα αποτελέσματα από τις προσομοιώσεις, όπως για παράδειγμα τα αποτελέσματα από το grid search, μπορούμε να χρησιμοποιήσου απευθείας το script plotGridSearch.m για το σχεδιασμό και την επίδειξη του πλέγματος.