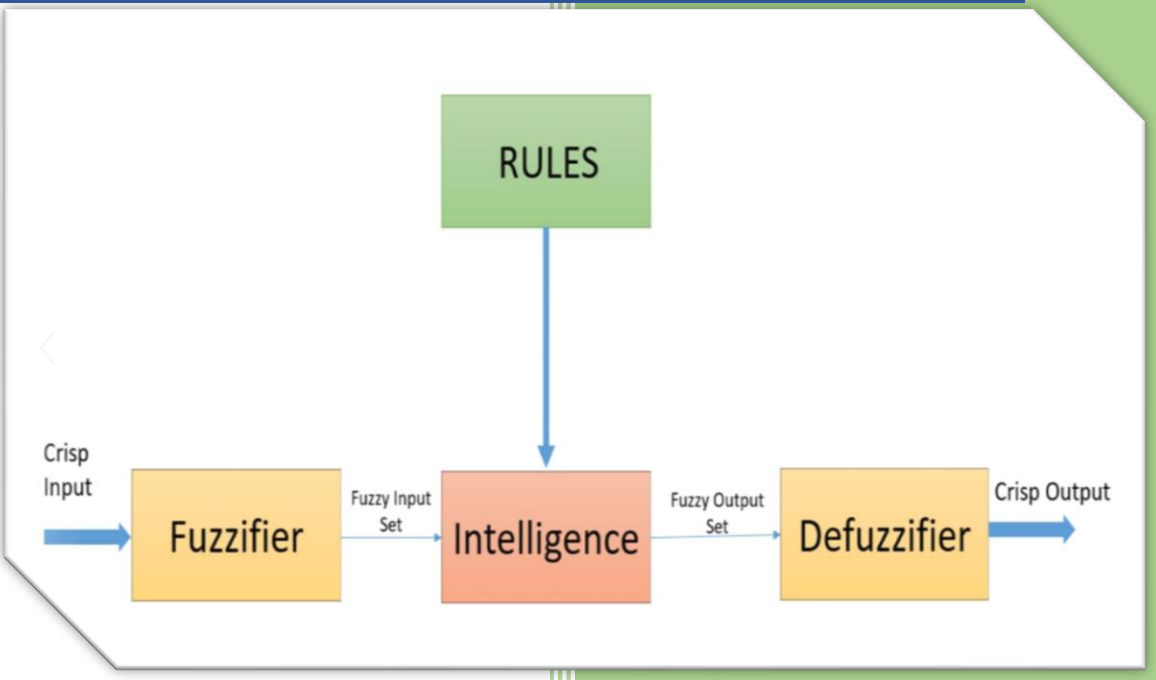
2019

Ασαφή Συστήματα – Εργασία ΙΙΙ



Τογκουσίδης Αναστάσιος

ΑΕΜ: 8920

01/8/2019

Πρώτο μέρος

Η τρίτη εργασία που μου ανατέθηκε είναι η **Group 3 Set 1**. Η εργασία διερευνά την ικανότητα των μοντέλων TSK να μοντελοποιούν μη γραμμικές συναρτήσεις πολλών μεταβλητών. Στο πρώτο μέρος της εργασίας επιλέγεται από το UCI Repository το Cobined Cycle Power Plant (CCPP) dataset το οποίο περιλαμβάνει 9568 instances και 4 features. Από εκφωνήσεις εργασιών παλαιότερων ετών που χρησιμοποιούσαν το εν λόγω data set φαίνεται πως τα δεδομένα συλλέχθηκαν από ένα εργοστάσιο παραγωγής ηλεκτρικής ενέργειας κατά τη διάρκεια 6 ετών και χαρακτηρίζονται από τις εξής τέσσερις μεταβλητές:

- Μέση ωριαία θερμοκρασία T
- Μέση ωριαία πίεση AP
- Μέση ωριαία σχετική υγρασία RH
- Μέση ωριαία αποβολή καυσαερίων V

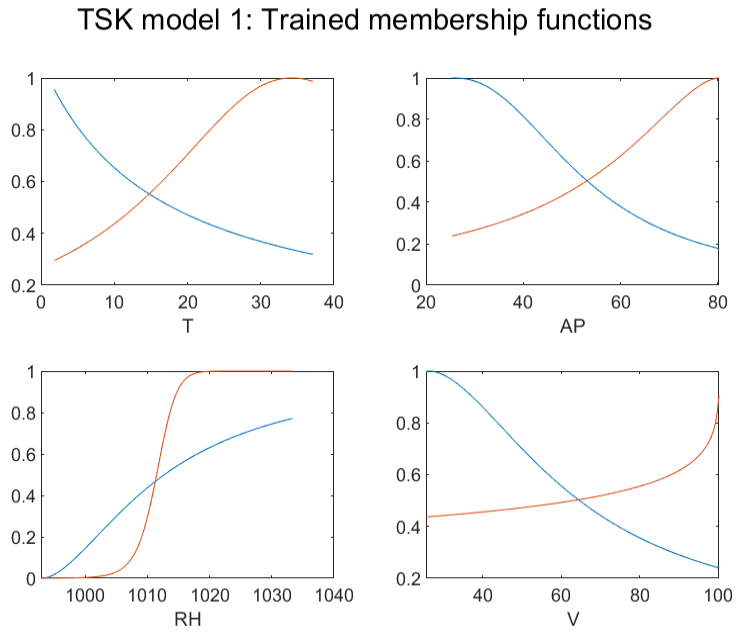
Η μεταβλητή που θεωρείται έξοδος στο σύστημα είναι η ενεργειακή απόδοση του σταθμού. Αρχικά, το σύνολο των δεδομένων διαχωρίζεται σε τρία, μη επικαλυπτόμενα υποσύνολα, το σύνολο εκπαίδευσης D_{trn} (60% των δειγμάτων), το σύνολο επικύρωσης D_{val} (20% των δειγμάτων) και το σύνολο ελέγχου D_{chk} (20% των δειγμάτων). Να σημειωθεί πως το αρχικό σύνολο των δειγμάτων αναδιατάσεται με τυχαίο τρόπο πριν διαμεριστεί στα επί μέρους σύνολα. Ζητείται να εκπαιδευτούν και να συγκριθούν 4 διαφοτερικά TSK μοντέλα που παρουσιάζονται 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

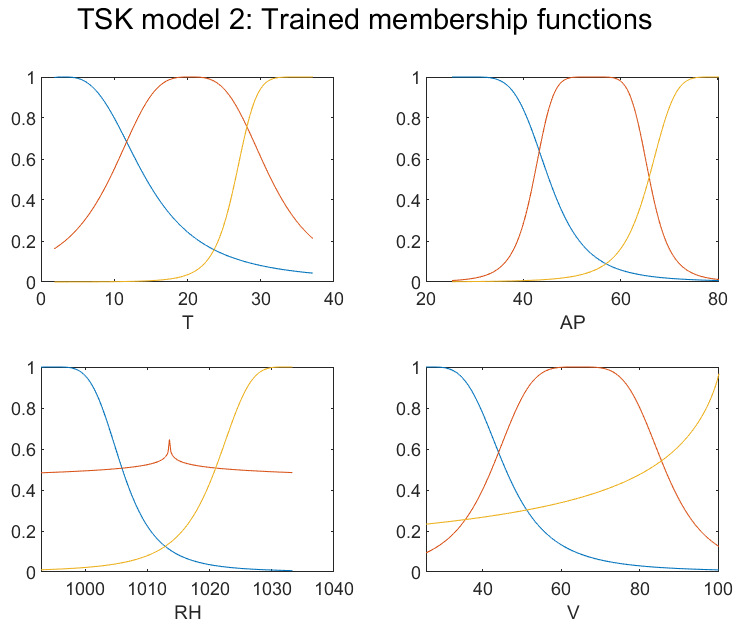
Πίνακας 1: Πίνακας μοντέλων που θα εκπεδευτούν

Για τη δημιουργία των μοντέλων χρησιμοποιείται η συνάρτηση `genfis1()` του Matlab. Ο λόγος που χρησιμοποιείται η συγκεκριμένη συνάρτηση είναι επειδή μας επιτρέπει να περάσουμε εύκολα τις παραμέτρους που θέλουμε στη δημιουργία του κάθε μοντέλου. Για την εκπαίδευση των μοντέλων χρησιμοποιείται η συνάρτηση `anfis()` του Matlab, η οποία χρησιμοποιεί την υβριδική μέθοδο, σύμφωνα με την οποία οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται μέσω της μεθόδου οπισθοδιάδοσης (backpropagation algorithm), ενώ οι παράμετροι της πολυωνυμικής συνάρτησης εξόδου βελτιστοποιούνται μέσω της μεθόδου ελαχίστων τετραγώνων (Least Squares Algorithm).

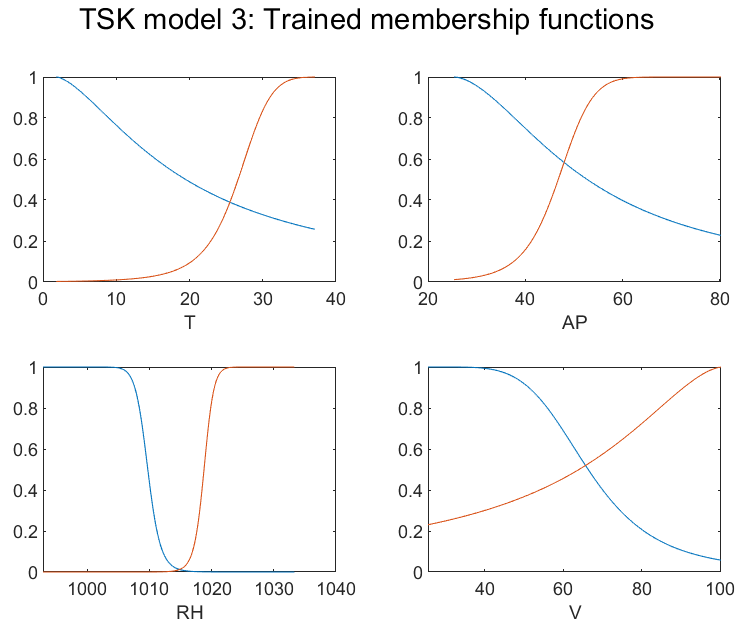
Στα σχήματα 1-4 παρουσιάζονται οι συναρτήσεις συμμετοχής των εκπαιδευμένων TSK μοντέλων:



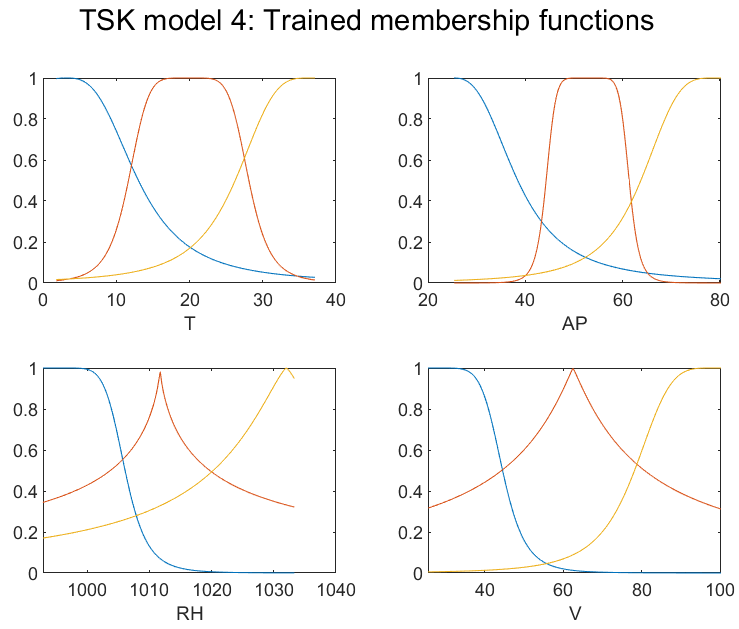
Σχήμα 1: Συναρτήσεις συμμετοχής εκπαιδευμένου μοντέλου TSK 1



Σχήμα 2: Συναρτήσεις συμμετοχής εκπαιδευμένου μοντέλου TSK 2

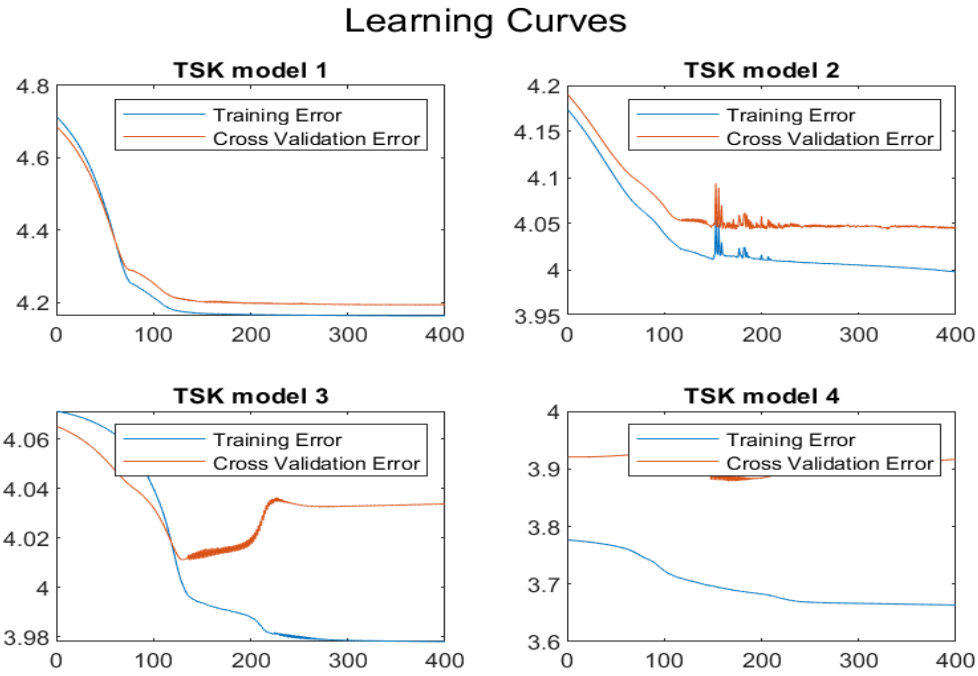


Σχήμα 3: Συναρτήσεις συμμετοχής εκπαιδευμένου μοντέλου TSK 3

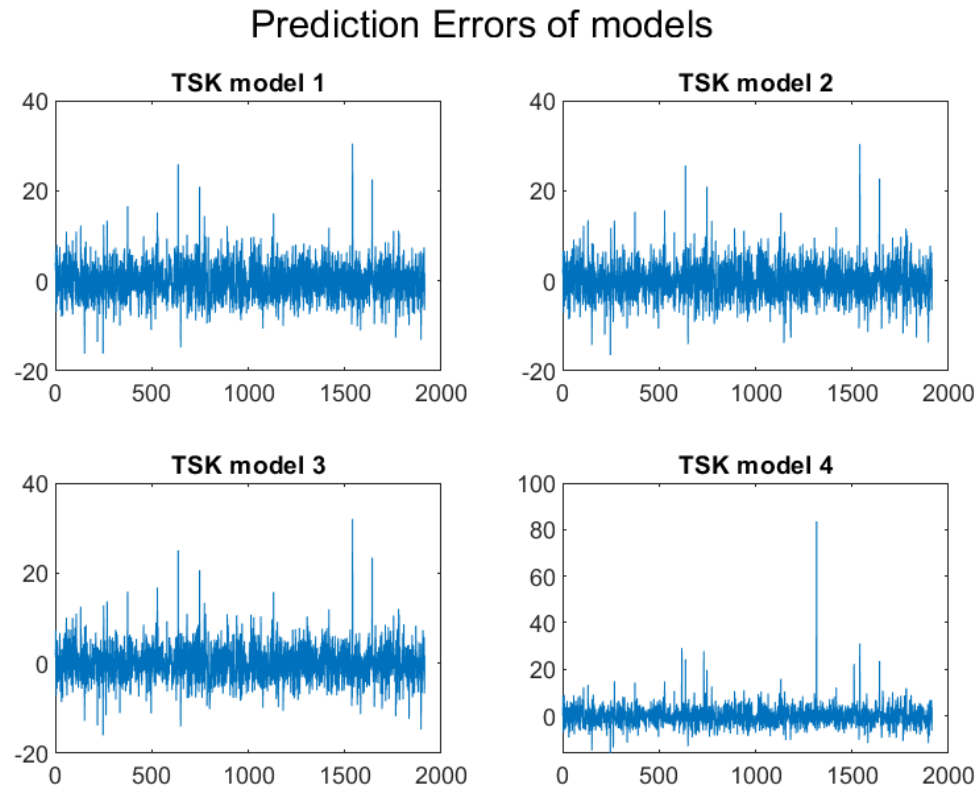


Σχήμα 4: Συναρτήσεις συμμετοχής εκπαιδευμένου μοντέλου TSK 4

Στο σχήμα 5 παρουσιάζονται οι καμπύλες εκμάθησης των τεσσάρων μοντέλων. Τέλος, στο σχήμα 6 παρουσιάζονται τα σφάλματα πρόβλεψης του κάθε μοντέλου ξεχωριστά.



Σχήμα 5: Καμπύλες εκμάθησης των τεσσάρων TSK μοντέλων



Σχήμα 6: Σφάλματα πρόβλεψης των τεσσάρων TSK μοντέλων

,

Στον πίνακα 2 παρουσιάζονται οι τιμές των δεικτών απόδοσης $RMSE$, $NMSE$, $NDEI$ και R^2 των τεσσάρων TSK μοντέλων.Οι δείκτες αυτοί υπολογίζονται με την εφαρμογή των εκπαιδευμένων πλέον μοντέλων στα test sets.

	$RMSE$	$NMSE$	$NDEI$	R^2
TSK model 1	4.2274	0.0605	0.2459	0.9395
TSK model 2	4.0923	0.0567	0.2381	0.9433
TSK model 3	4.0690	0.0560	0.2367	0.9440
TSK model 4	4.4882	0.0682	0.2611	0.9318

Πίνακας 2: Πίνακας δεικτών απόδοσης των TSK μοντέλων

Το φαινόμενο της υπερεκπαίδευσης παρατηρείται στα TSK μοντέλα 3 και 4. Συγκεκριμένα, παρ'όλο που το σφάλμα εκπαίδευσης και στα δύο αυτά αυτά μοντέλα φαίνεται να μειώνεται σε κάθε επανάληψη, το σφάλμα στο σετ αξιολόγησης παραμένει σχεδόν σταθερό. Σύμφωνα με τα αποτελέσματά μου, το φαινόμενο της υπερεκπαίδευσης οφείλεται περισσότερο στη μορφή της εξόδου (πολυωνυμική) παρά στον αριθμό των συναρτήσεων συμμετοχής.

Δεύτερο Μέρος

Το δεύτερο μέρος της εργασίας πραγματεύεται την εφαρμογή μια πιο συστηματική προσέγγιση στο πρόβλημα μοντελοποίησης μιας άγνωστης συνάρτησης σε dataset με υψηλή διαστασιμότητα. Η ιδέα είναι ότι ο αλγόριθμος αναζήτησης πλέγματος (grid search) και αξιολόγησης μέσω πεντάπτυχης διασταυρωμένης επικύρωσης (5-fold cross validation) προσπαθεί να βρει τις βέλτιστες τιμές στους δύο βαθμούς ελευθερίας που έχει το πρόβλημα, δηλαδή των αριθμό των χαρακτηριστικών που θα ληφθούν υπ' όψιν και τον τον αριθμό των συναρτήσεων συμμετοχής.

Αρχικά, εφαρμόζεται ο αλγόριθμος Relief ώστε να τοποθετηθούν τα features των δειγμάτων στη βέλτιστη σειρά. Σε κάθε επανάληψη της πεντάπτυχους διασταυρωμένης επικύρωσης γίνεται τυχαία αναδιάταξη των δεδομένων και διαχωρισμός τους και training set (80% των δεδομένων) και test set (20% των δεδομένων) με τη βοήθεια της συνάρτησης `cvpartition()` του Matlab. Για τη δημιουργία του μοντέλου χρησιμοποιείται η συνάρτηση `genfis3()` του Matlab, η οποία χρησιμοποιεί τον αλγόριθμο Fuzzy C-means για την ομαδοποίηση των δεδομένων. Το μοντέλο εκπαιδεύεται με τη συνάρτηση `anfis()` του Matlab και τα αποτελέσματα των μέσων όρων των σφαλμάτων σε κάθε περίπτωση συσσωρεύονται στον πίνακα 3.

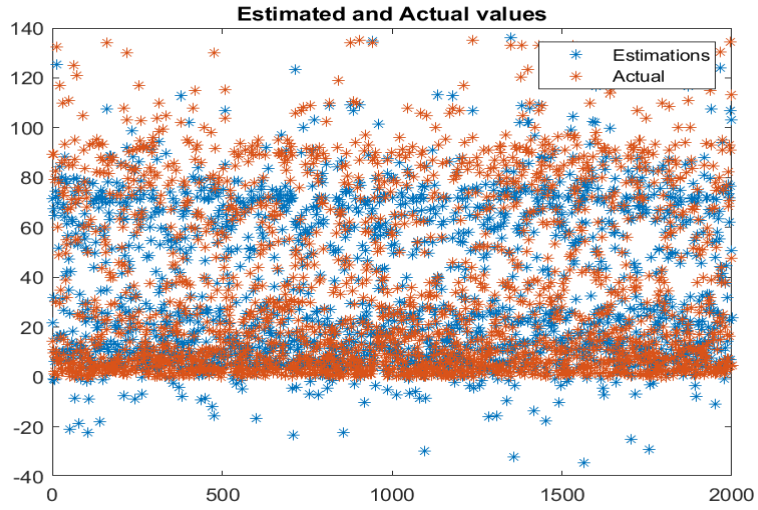
NF / NR	3	6	9	12
3	21.3729	20.0464	19.9956	17.6465
6	20.7828	19.4576	18.4441	17.2564
9	20.1496	18.7601	17.6631	16.9698
12	19.7705	18.5147	17.1003	16.5736
15	19.9875	18.4253	16.9945	16.1693

Πίνακας 3: Μέσο σφάλμα πεντάπτυχους διασταυρωμένης επικύρωσης

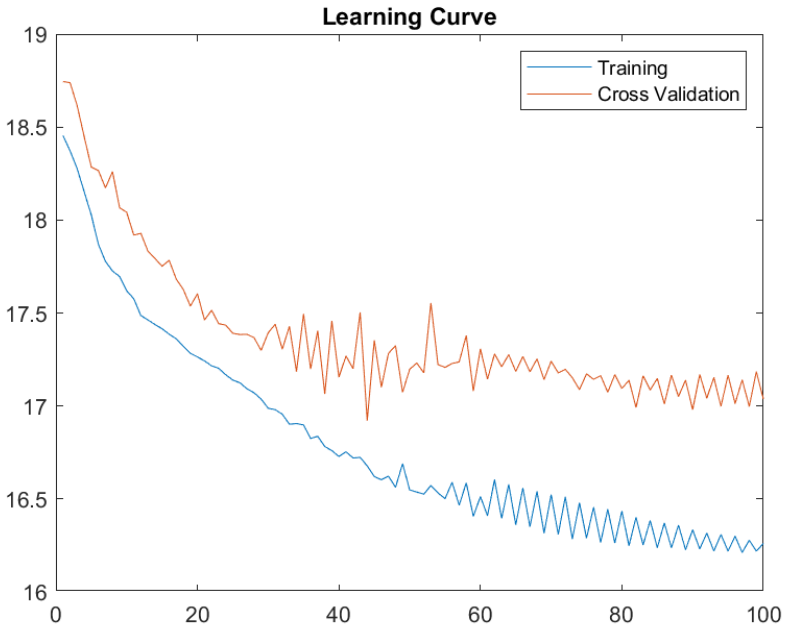
Η βέλτιστη τιμή του παραπάνω πίνακα μπορεί να δίνει την αίσθηση ότι είναι μεγάλη τιμή, ωστόσο πρέπει να ληφθεί υπ' όψιν ότι οι τιμές εξόδου κυμαίνονται περίπου στο διάστημα $[0,160]$. Συνεπώς, το μέσο σφάλμα πενταπτυχούς διασταυρωμένης επικύρωσης δίνει μια μικρή απόκλιση $\mathbf{8}\%$ στις προβλέψεις.

Να σημειωθεί πως **λόγω χρονικών περιορισμών** (χρειαζόμουν το λάπτοπ μου για τη διπλωματική εργασία) επιλέχθηκαν μικρότερες τιμές στα μεγέθη NR και NF, προκειμένου να μειωθεί ο χρόνος της διαδικασίας εκπαίδευσης. Επιπλέον δούλεψα με τον μισό αριθμό δεδομένων (10.000).

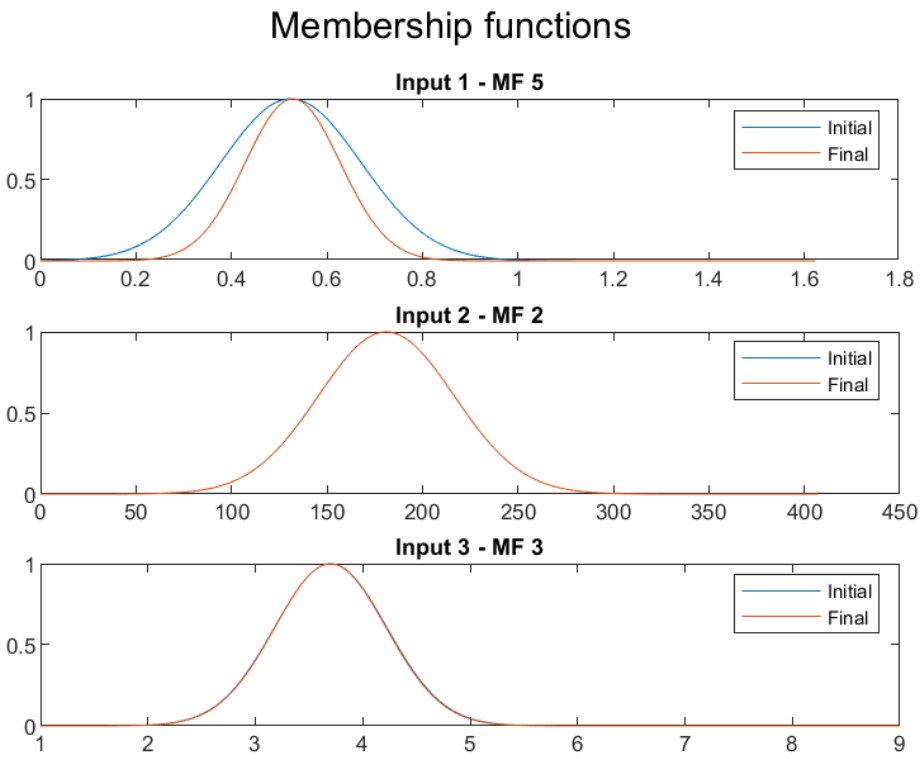
Επιλέγονται, λοιπόν, ως βέλτιστες επιλογές οι $\boldsymbol{NF = 12}$ και $\boldsymbol{NR = 15}$. Στη συνέχεια, γίνεται ξανά εκπαίδευση του μοντέλου με τα βέλτιστα χαρακτηριστικά. Τα διαγράμματα που προκύπτουν παρουσιάζονται στα σχήματα 7 έως 9.



Σχήμα 7: Εκτιμήσεις και πραγματικές τιμές εξόδου



Σχήμα 8: Καμπύλη εκμάθησης μοντέλου



Σχήμα 9: Παραδείγματα συναρτήσεων συμμετοχής πριν και μετά την εκμάθηση

Να σημειωθεί πως οι συναρτήσεις συμμετοχής που απεικονίζονται παραπάνω επιλεχθηκαν τυχαία. Τέλος, οι ζητούμενες μετρικές σφάλματος που προκύπτουν από την παραπάνω διαδιασία είναι:

$$RMSE = 15.5150$$

$$NMSE = 0.2064$$

$$NDEI = 0.4543$$

$$R^2 = 0.7936$$

Αν είχαμε επιλέξει grid partition με δύο ή τρία ασαφή σύνολα στην είσοδο θα είχαμε 2^{81} και 3^{81} κανόνες αντίστοιχα. Η πολυπλοκότητα του μοντέλου θα ήταν τελείως ασύμορφη και η υλοποίηση πρακτικά αδύνατη!