1. Εφαρμογή σε απλό dataset

1.1. Κώδικας 1ου μέρους εργασίας

Ο κώδικας σε MATLAB που υλοποιεί το μέρος α' της εργασίας βρίσκεται στο αρχείο */3/matlab/main_a.m*. Εκεί υπάρχει μόνο η λογική της εκτέλεσης, ενώ βοηθητικές κλάσεις και συναρτήσεις υπάρχουν στο φάκελο */Matlab Helpers/**.

1.2. Φόρτωση & Προ-επεξεργασία dataset

Το dataset του ερωτήματος, *CCPP*, αποτελείται από 9568 δείγματα (data points) με (4) features και μία τιμή εξόδου το καθένα. Δεν υλοποιήθηκε κάποια μέθοδος προ-επεξεργασίας του dataset καθώς τα αποτελέσματα ήταν καλά. Το μόνο που γίνεται μετά την φόρτωση είναι ο έλεγχος για διπλότυπα δείγματα (~40 δείγματα αφαιρέθηκαν).

1.3. Διαχωρισμός του dataset

Για το διαχωρισμό του dataset στα **training** (χρησ. στο training), **validation** (χρησ. στο training για αποφυγή overfitting) και **testing** (χρησ. στο testing – άγνωστο apriori) subsets δημιουργήθηκε μια μέθοδος, *AnfisWrapper.partition()*, η οποία δέχεται σαν ορίσματα το dataset και τα ποσοστά του διαχωρισμού και επιστρέφει τα τρία subsets παραπάνω. Ο τρόπος που υλοποιεί το splitting είναι μία τύπου round-robin ανάθεση στοιχείων στα subsets βάσει των ποσοστών. Για παράδειγμα, για splitting 60% - 20% - 20% η μεταβλητή *split* θα είναι *[0.6, 0.2, 0.2]*, ενώ η ανάθεση θα γίνει ως εξής:

```
split10 = 10 * split = [6, 2, 2];
Για κάθε 10άδα στοιχείων του dataset, do:
      Βάλε 6 στο training
      Βάλε 2 στο validation
      Βάλε 2 στο testing
end
```

Έτσι προκύπτει μια δίκαια ανάθεση με τον τελικό αριθμό στοιχείων σε κάθε ένα από τα 3 subsets να *x * |dataset| ± 1*, με x το ποσοστό.

1.4. Ζητούμενα Εργασίας

Ζητείται να εκπαιδευθούν τέσσερα (4) TSK μοντέλα για το παραπάνω dataset με (4) εισόδους το καθένα και πλήθος MFs για κάθε είσοδο και τύπο εξόδου του sugeno μοντέλου, που καθορίζεται ως εξής:

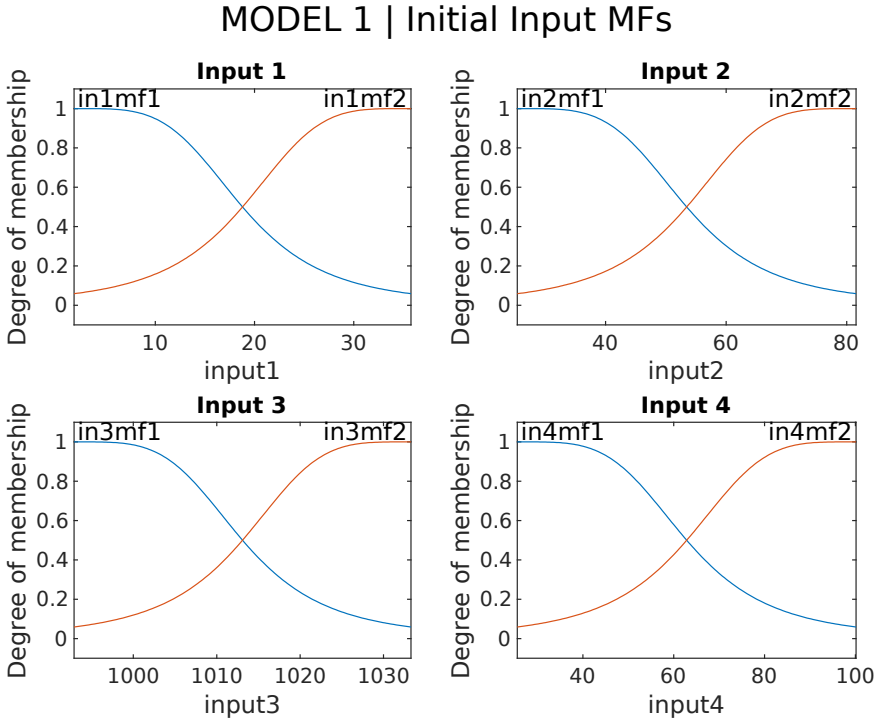
	Πλήθος MFs Μεταβλητών Εισόδου	Τύπος Εξόδου Κανόνων Sugeno
TSK Model 1	2	Singleton
TSK Model 2	3	Singleton
TSK Model 3	2	Linear
TSK Model 4	3	Linear

Πιν. 1: Παράμετροι των TSK μοντέλων που θα εκπαιδευτούν

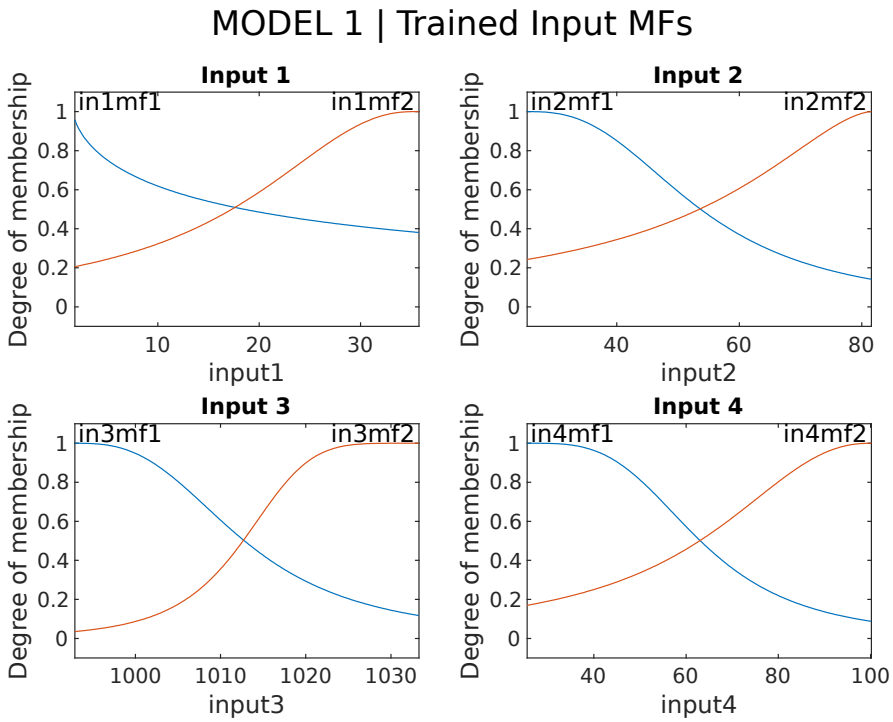
1.5. Εκπαίδευση TSK μοντέλων

1.5.1 Model 1: (2) Input MFs, Singleton Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

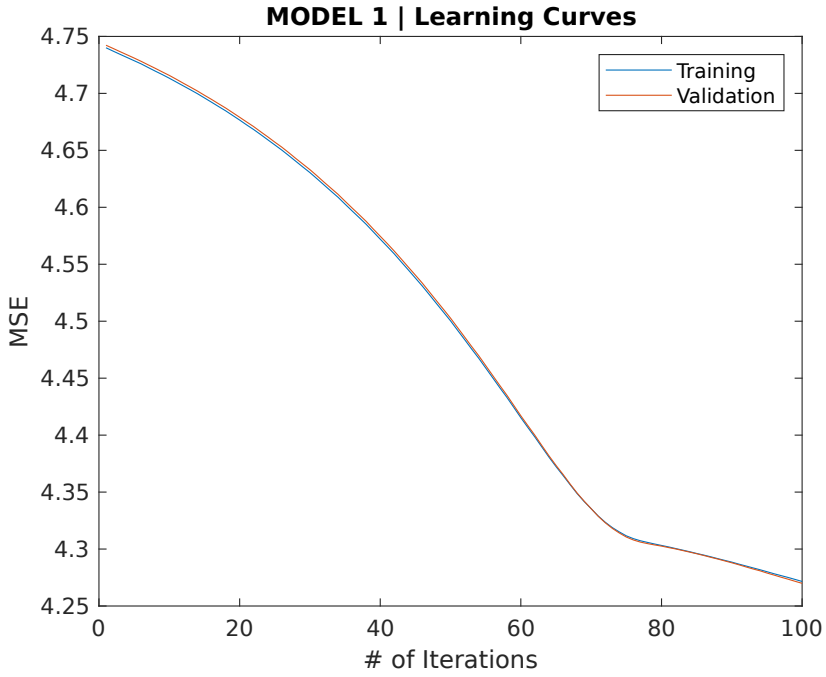


Εικ. 1: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 1ου μοντέλου

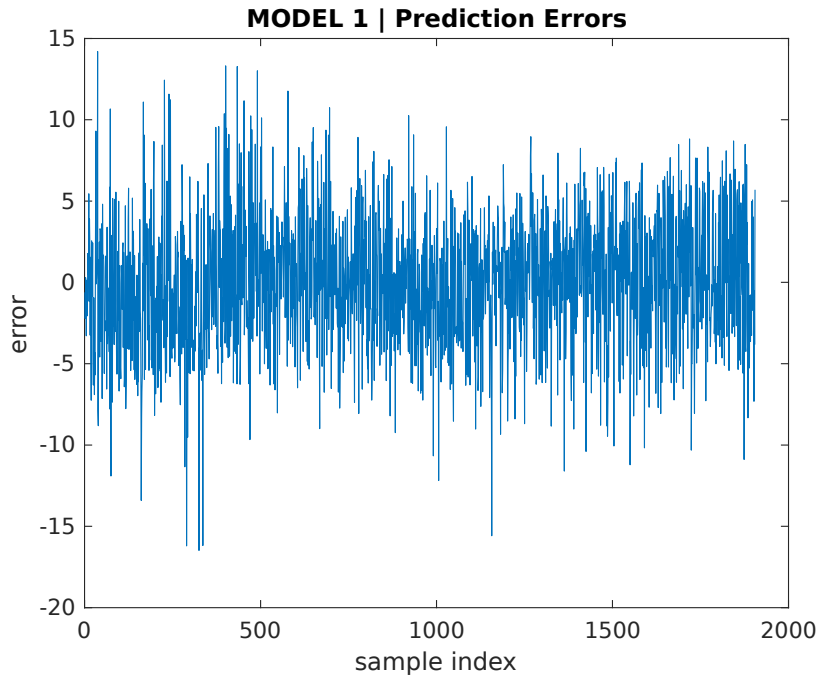


Εικ. 2: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 1ου μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικ. 3: Καμπύλες μάθησης 1ου μοντέλου



Εικ. 4: Σφάλματα πρόβλεψης κατά την εφαρμογή του 1ου μοντέλου στο test set

Τέλος, δίνεται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	**MSE**	**RMSE**	**NMSE**	**R^2**	**NDEI**
Value →	16.627	4.078	0.062	0.938 (**93.8%**)	0.249

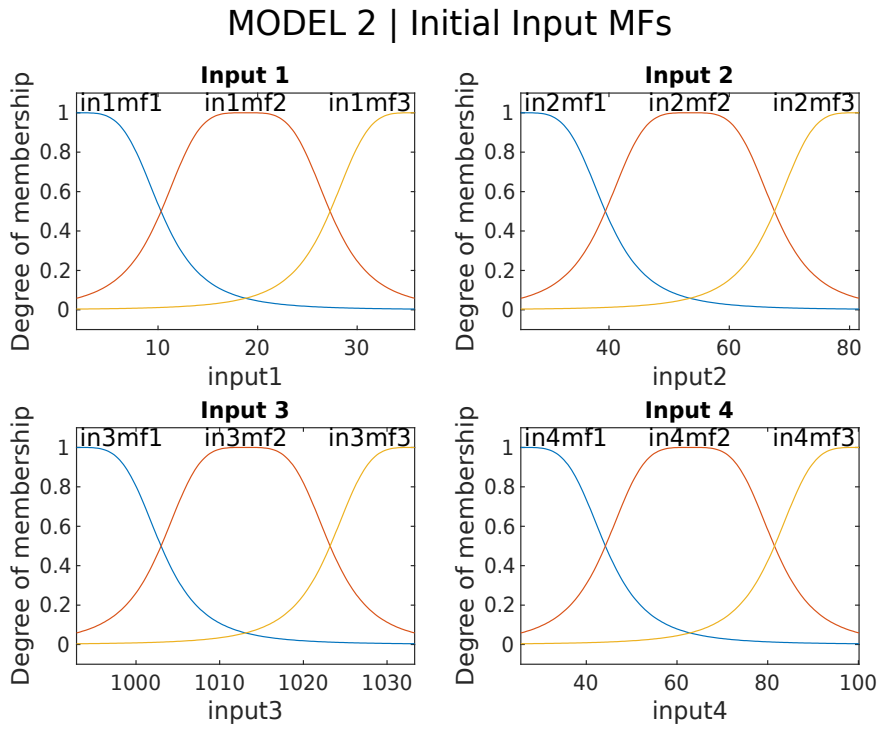
Πιν. 2: Μετρικές Απόδοσης του 1ου μοντέλου

Μια πρώτη ανάλυση:

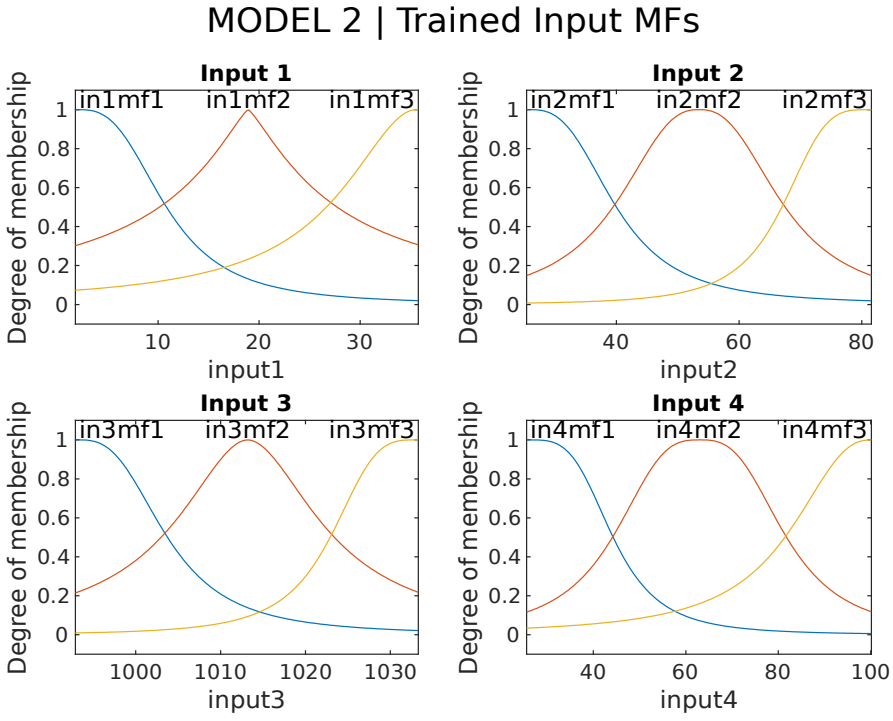
Το εκπαιδευμένο μοντέλο απέδωσε καλά στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι περίπου 94%. Επίσης όπως φαίνεται από τα learning curves δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***). Τα παραπάνω επιβεβαιώνονται και στο διάγραμμα των prediction errors που βρίσκονται γύρω από το και με σχετικά μικρή για τα μεγέθη της εξόδου διακύμανση.

1.5.2 Model 2: (3) Input MFs, Singleton Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

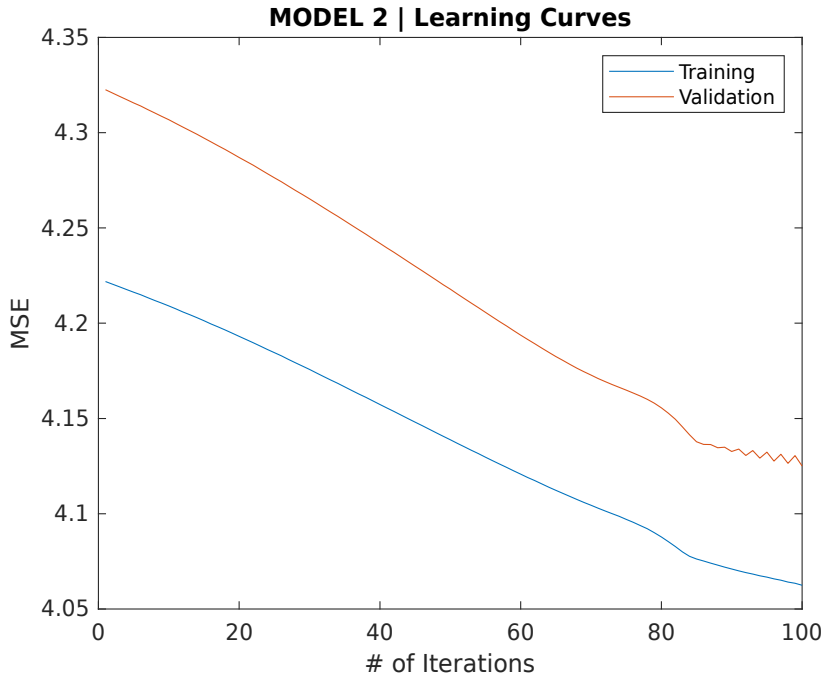


Εικ. 5: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 2ου μοντέλου

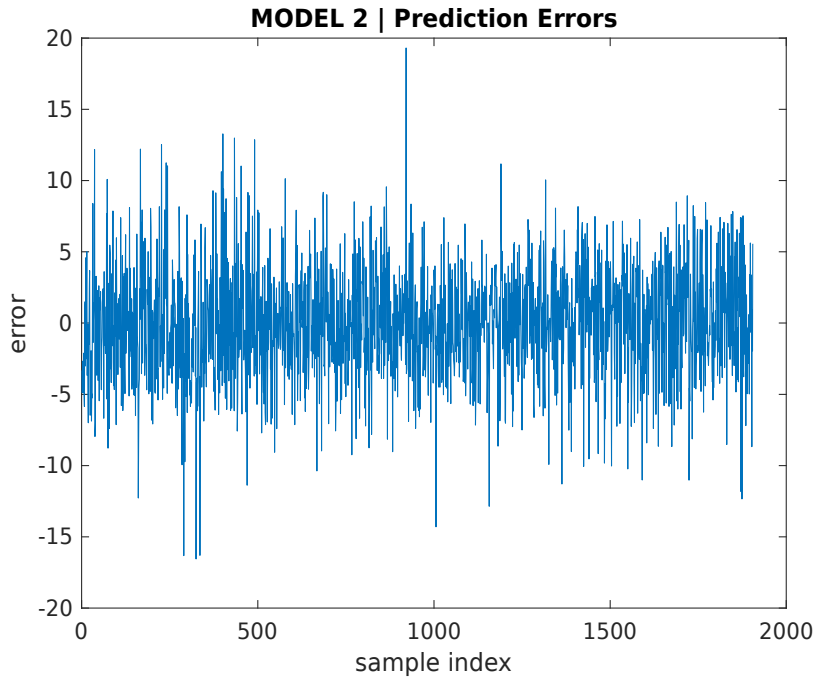


Εικ. 6: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 2ου μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικ. 7: Καμπύλες μάθησης 2ου μοντέλου



Εικ. 8: Σφάλματα πρόβλεψης κατά την εφαρμογή του 2ου μοντέλου στο test set

Τέλος, δίνεται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	**MSE**	**RMSE**	**NMSE**	**R^2**	**NDEI**
Value →	15.679	3.96	0.058	0.9423 (**94.23%**)	0.24

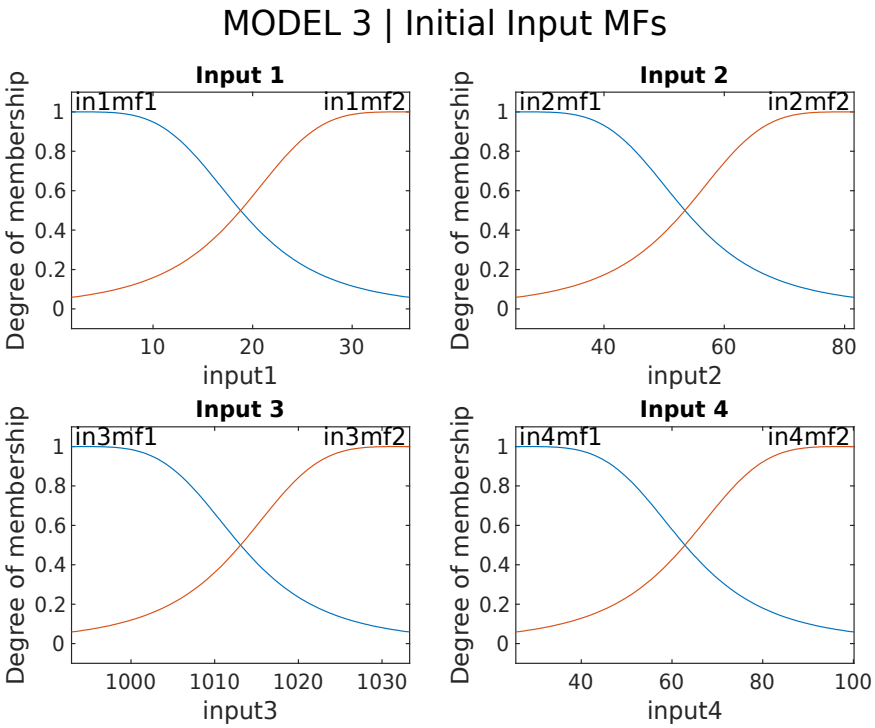
Πιν. 3: Μετρικές Απόδοσης του 2ου μοντέλου

Μια πρώτη ανάλυση:

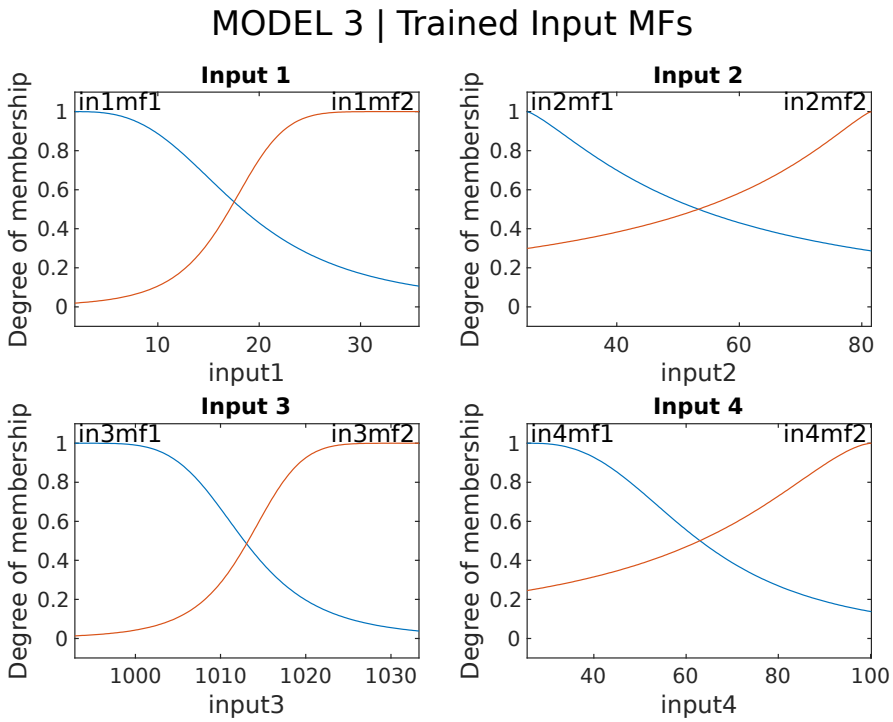
Το δεύτερο εκπαιδευμένο μοντέλο απέδωσε λίγο καλύτερα από το πρώτο στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι λίγο μεγαλύτερος από 94%. Επίσης όπως φαίνεται από τα learning curves δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***). Τα παραπάνω επιβεβαιώνονται και στο διάγραμμα των prediction errors που βρίσκονται γύρω από το και με σχετικά μικρή για τα μεγέθη της εξόδου διακύμανση.

1.5.3 Model 3: (2) Input MFs, Polynomial Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

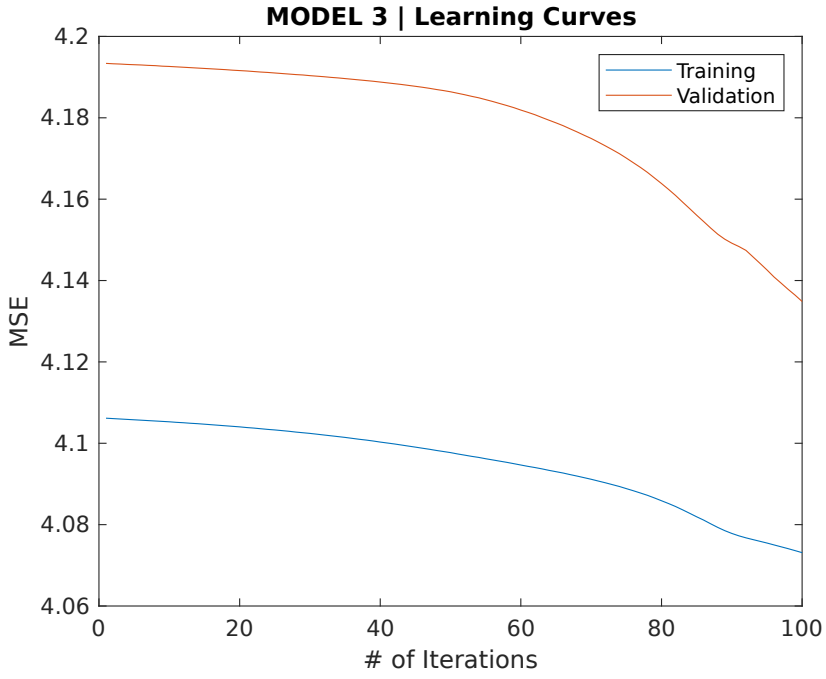


Εικ. 9: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 3ου μοντέλου

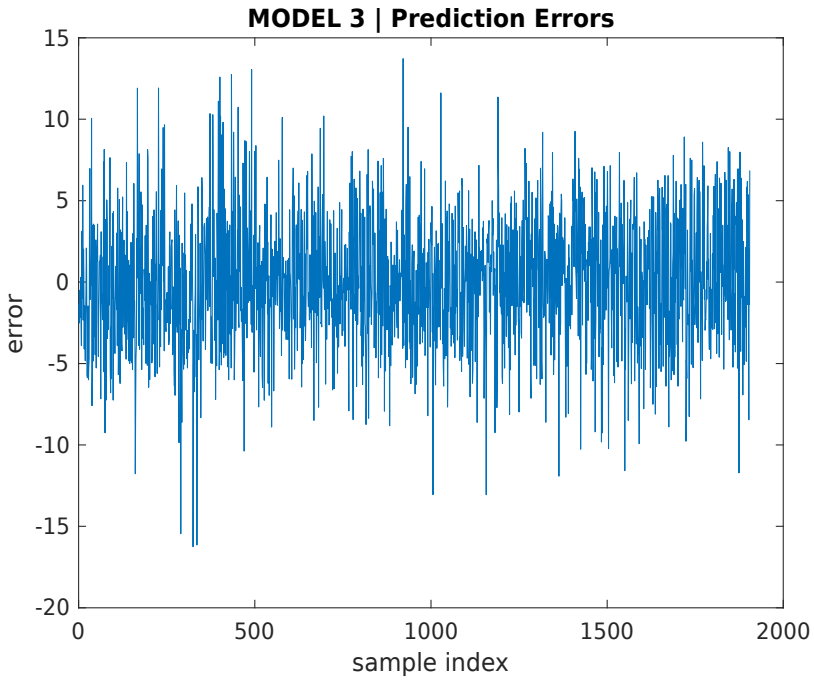


Εικ. 10: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 3ου μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικ. 11: Καμπύλες μάθησης 3ου μοντέλου



Εικ. 12: Σφάλματα πρόβλεψης κατά την εφαρμογή του 3ου μοντέλου στο test set

Τέλος, δίνεται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	**MSE**	**RMSE**	**NMSE**	**R^2**	**NDEI**
Value →	15.364	3.92	0.057	0.9433 (**94.33%**)	0.238

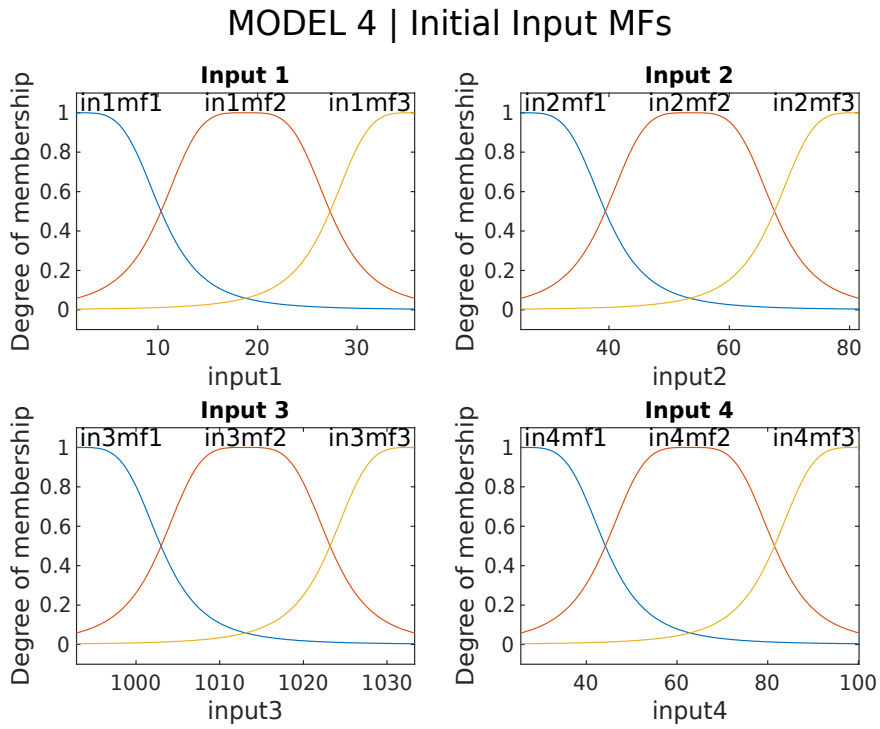
Πιν. 4: Μετρικές Απόδοσης του 3ου μοντέλου

Μια πρώτη ανάλυση:

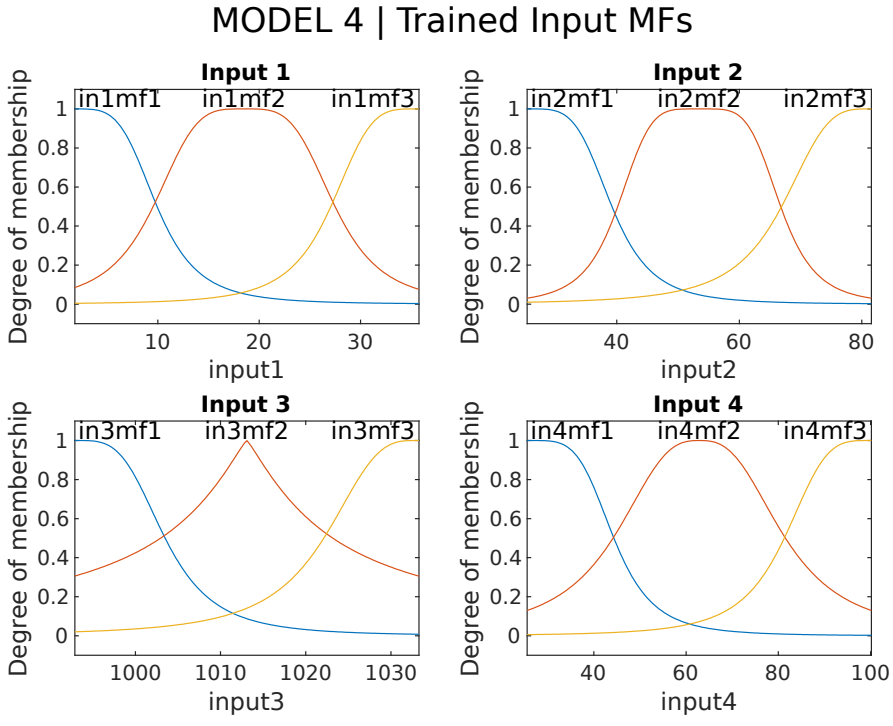
Το τρίτο εκπαιδευμένο μοντέλο απέδωσε λίγο καλύτερα από το πρώτο (έχουν ίδιο αριθμό MFs των μεταβλητών εισόδου) στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι 94.33% (vs. 93.8%). Επίσης όπως φαίνεται από τα learning curves δεν υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***). Τα παραπάνω επιβεβαιώνονται και στο διάγραμμα των prediction errors που βρίσκονται γύρω από το και με σχετικά μικρή για τα μεγέθη της εξόδου διακύμανση.

1.5.4 Model 4: (3) Input MFs, Polynomial Output

Αρχικά, παρατίθενται οι αρχικές και τελικές μορφές των MFs για τις ασαφείς τιμές των ασαφών μεταβλητών εισόδου των κανόνων:

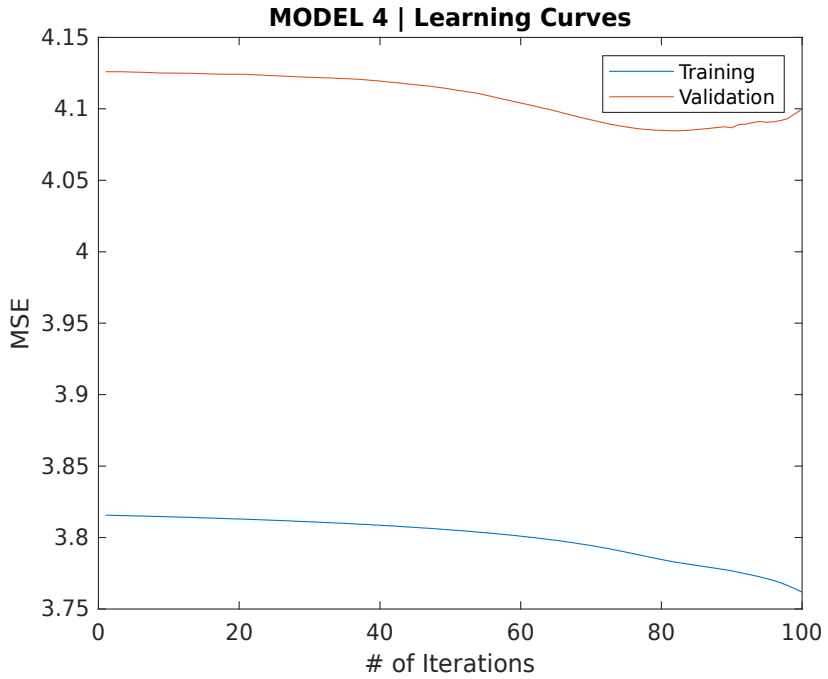


Εικ. 13: Αρχική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 4ου μοντέλου

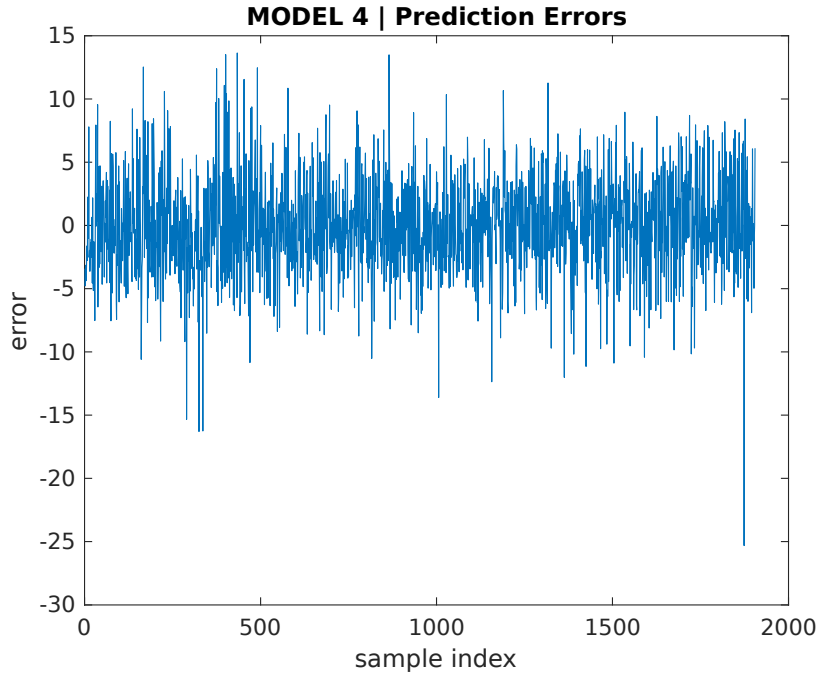


Εικ. 14: Τελική μορφή συναρτήσεων συμμετοχής μεταβλητών εισόδου 4ου μοντέλου

Ακολούθως, δίνονται οι καμπύλες μάθησης (training) του μοντέλου καθώς και τα σφάλματα κατά την εφαρμογή του εκπαιδευμένου μοντέλου (testing) στο test set:



Εικ. 15: Καμπύλες μάθησης 4ου μοντέλου



Εικ. 16: Σφάλματα πρόβλεψης κατά την εφαρμογή του 4ου μοντέλου στο test set

Τέλος, δίνεται οι ζητούμενες μετρικές απόδοσης του μοντέλου:

Metric →	**MSE**	**RMSE**	**NMSE**	**R^2**	**NDEI**
Value →	14.848	3.853	0.054	0.946 (**94.6%**)	0.232

Πιν. 5: Μετρικές Απόδοσης του 4ου μοντέλου

Μια πρώτη ανάλυση:

Το τέταρτο μοντέλο από αυτά που εκπαιδεύτηκαν απέδωσε λίγο καλύτερα από το δεύτερο (έχουν ίδιο αριθμό MFs των μεταβλητών εισόδου) στο regression task στο testing subset με τον δείκτη απόδοσης R^2 να είναι 94.6% (vs. 94.23%). Όμως, όπως φαίνεται από τα learning curves υπάρχει το φαινόμενο της υπερεκπαίδευσης (overfitting – ***epochNumber=100***) για το ίδιο epochNumber με τα υπόλοιπα. Να τονισθεί ότι, να και είναι λίγο παράδοξο, για 80 epochs το παραπάνω μοντέλο ενώ δεν παρουσιάζει overfitting δεν υπάρχει καμία βελτίωση της απόδοσης (μάλιστα υπάρχει και μία μικρή χειροτέρευση – 94.59%). Γι' αυτό δεν παραθέτονται τα στοιχεία του 4ου μοντέλου χωρίς overfitting.

Συμπερασματική ανάλυση:

Αρχικά παραθέτονται συγκεντρωτικά οι μετρικές απόδοσης όλων των μοντέλων:

Metric → ↓ Model	**MSE**	**RMSE**	**NMSE**	**R^2**	**NDEI**
Model 1	16.627	4.078	0.062	0.938 (**93.80%**)	0.249
Model 2	15.679	3.960	0.058	0.942 (**94.23%**)	0.240
Model 3	15.364	3.920	0.057	0.943 (**94.33%**)	0.238
Model 4	14.848	3.853	0.054	0.946 (**94.6%**)	0.232

Πιν. 6: Συγκεντρωτικές Μετρικές Απόδοσης όλων των μοντέλων

Βάσει των παραπάνω μετρικών απόδοσης φαίνεται ότι:

1. Αύξηση του πλήθους των συναρτήσεων συμμετοχής (MFs) και άρα των πιθανών τιμών ανά ασαφή μεταβλητή εισόδου οδηγεί σε καλύτερα αποτελέσματα για ίδια μορφή εξόδου του μοντέλου (διαφορά μοντέλου 1 από 2, διαφορά μοντέλου 3 από 4)
2. Για ίδιο πλήθος συναρτήσεων συμμετοχής (MFs) ανά ασαφή μεταβλητή εισόδου, διατήρηση περισσότερων όρων στην έξοδο του κάθε κανόνα του μοντέλου sugeno (μετάβαση από σταθερή -singleton- έξοδο σε πολυωνυμική –polynomial-) οδηγεί σε καλύτερα αποτελέσματα (διαφορά μοντέλου 1 από 3, διαφορά μοντέλου 2 από 4)

3. Όλα τα μοντέλα, όπως έχει αναφερθεί και στις πρώτες αναλύσεις, παρουσιάζουν αρκετά ικανοποιητική απόδοση παλινδρόμησης (regression) με τον δείκτη απόδοσης R2 να κυμαίνεται περίπου μεταξύ 94% και 95%. Ωστόσο το **μοντέλο 4** επιλέγεται ως το πιο αποδοτικό με βάση τον παραπάνω πίνακα.
4. Αξίζει να σημειωθεί ότι το μοντέλο (2) (3 συναρτήσεις συμμετοχής, singleton έξοδος) είναι πολύ κοντά σε απόδοση από το μοντέλο 3 (2 συναρτήσεις συμμετοχής, polynomial έξοδος) το οποίο εν μέρη είναι αναμενόμενο.
5. Τέλος, στα πρώτα τρία μοντέλα δεν παρατηρείται overfitting σε αντίθεση με το τέταρτο, γεγονός που φαίνεται από την απόκλιση της καμπύλης μάθησης στο validation set, η όποια δεν εμφανίζει μονότονα φθίνουσα μορφή (όπως εμφανίζει στο training set). Ο αριθμός των epochs (επαναλήψεων) στην εκπαίδευση του κάθε μοντέλου (FNN) είναι 100.

2. Εφαρμογή σε high-dimensional dataset

2.1. Κώδικας 2ου μέρους εργασίας

Ο κώδικας σε MATLAB που υλοποιεί το μέρος α' της εργασίας βρίσκεται στο αρχείο */3/matlab/main_b.m*. Εκεί υπάρχει μόνο η λογική της εκτέλεσης, ενώ βοηθητικές κλάσεις και συναρτήσεις υπάρχουν στο φάκελο */Matlab Helpers/**.