Εργασία 3 - Υπολογιστική Νοημοσύνη

Επίλυση προβλήματος παλινδρόμησης με χρήση μοντέλων TSK

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Εργασία του: **Παπαδόπουλου Κωνσταντίνου**

ΑΕΜ: **8677**

Εισαγωγή

Στόχος της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των μοντέλων TSK στη μοντελοποίηση πολυμεταβλητών, μη γραμμικών συναρτήσεων, σύμφωνα με τα ζητούμενα της εκφώνησης της Εργασίας 3 του μαθήματος Υπολογιστικής Νοημοσύνης.

Στη συνέχεια, αναλύουμε τη διαδικασία εργασίας πάνω στα ζητούμενα προβλήματα, σχολιάζοντας και τη λογική συγγραφής του κώδικα, όπου αυτό κρίνεται απαραίτητο (υπάρχει σχολιασμός και στον ίδιο τον πηγαίο κώδικα, ο οποίος είναι σε Matlab) και παραθέτουμε τα απαραίτητα διαγράμματα.

1 Εφαρμογή σε απλό dataset

Στην πρώτη φάση της εργασίας, επιλέγεται από το UCI repository το Airfoil Self-Noise dataset[1], το οποίο περιλαμβάνει 1503 δείγματα (instances) και 6 χαρακτηριστικά (features) και συγκεκριμένα 5 χαρακτηριστικά δεδομένων και 1 χαρακτηριστικό εξόδου (αυτό της τελευταίας στήλης).

Στη συνέχεια ακολουθούμε τα εξής βήματα:

- ✓ *Διαχωρίζουμε τα δεδομένα σε σύνολα εκπαίδευσης-επικύρωσης-ελέγχου*
  Αρχικά, ελέγχουμε ότι στο dataset δεν υπάρχουν missing values, επομένως ούτε ανάγκη διαχείρισης αυτών.

[1] https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise

Βλέπουμε ότι στο *Abstract: Missing Values?-N/A,* αλλά περνώντας το dataset διαπιστώνουμε ότι μπορούμε ορθώς να υποθέσουμε ότι δεν έχουμε missing values.

Πέρα από το διαχωρισμό των 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 μοντέλου, προβαίνουμε επίσης σε μία διαδικασία προεπεξεργασίας των δεδομένων μας (βλ. *preprocess_samples.m*).

Συγκεκριμένα, δεν «σπάμε» απλώς το σύνολο των δεδομένων στα πρώτα 60%, 20%, 20% δείγματα, αλλά επιδιώκουμε τα δείγματα σε κάθε σύνολο να είναι αρκετά αντιπροσωπευτικά και να περιλαμβάνουν στατιστικά παρόμοια πληροφορία σε σχέση με το σύστημα. Επειδή λοιπόν παρατηρώντας το datasheet τα χαρακτηριστικά ορισμένων στηλών φαίνονται να είναι ομαδοποιημένα, σπάμε τα σύνολά μας με ένα τυχαίο τρόπο (χρησιμοποιώντας και τη συνάρτηση *datasample()* της Matlab), ο οποίος φαίνεται αναλυτικά και στον πηγαίο κώδικα.

Επιπλέον, φροντίζουμε να συμπεριλάβουμε τις μέγιστες και τις ελάχιστες τιμές του κάθε χαρακτηριστικού στο σύνολο εκπαίδευσης, έτσι ώστε να έχουμε ένα ακόμη πιο αντιπροσωπευτικό σύνολο σε εύρος.

- ✓ *Εκπαιδεύουμε TSK μοντέλα με διαφορετικές παραμέτρους*

  Σε αυτό το στάδιο θα εξεταστούν τα 4 μοντέλα TSK που δίνονται στην εκφώνηση, τα οποία διαφέρουν προς τη μορφή της εξόδου τους και το πλήθος των συναρτήσεων συμμετοχής.

  Τα μοντέλα θα δημιουργηθούν χρησιμοποιώντας τη συνάρτηση *genfis1()* στο Matlab, με την οποία μπορούμε να σχηματίσουμε το

grid partitioning που θέλουμε, επιλέγοντας τη μορφή της συνάρτησης συμμετοχής (σε αυτήν την περίπτωση bell-shaped) και με τον βαθμό επικάλυψης που επιθυμούμε (0.5, το οποίο είναι και το default).

Για την εκπαίδευση των μοντέλων χρησιμοποιούμε τη συνάρτηση *anfis()*, η οποία χρησιμοποιεί 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).

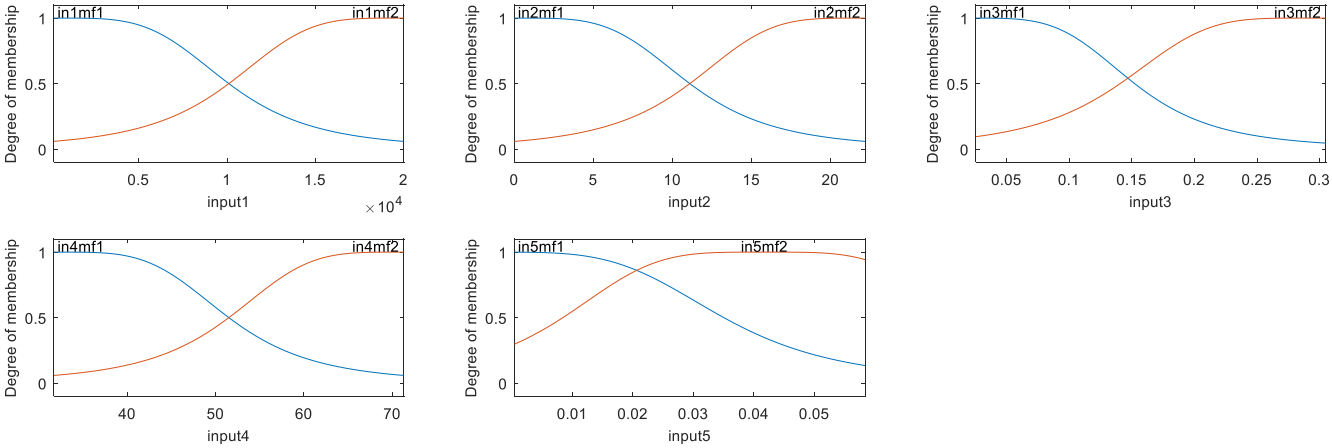
- ✓ *Καθορίζουμε τους δείκτες απόδοσης για την αξιολόγηση των μοντέλων*

  Στο αρχείο *evaluate.m* γράφουμε τις υλοποιήσεις των μετρικών που θα χρησιμοποιήσουμε, δηλαδή των MSE, RMSE, NMSE, NDEI και R^2.

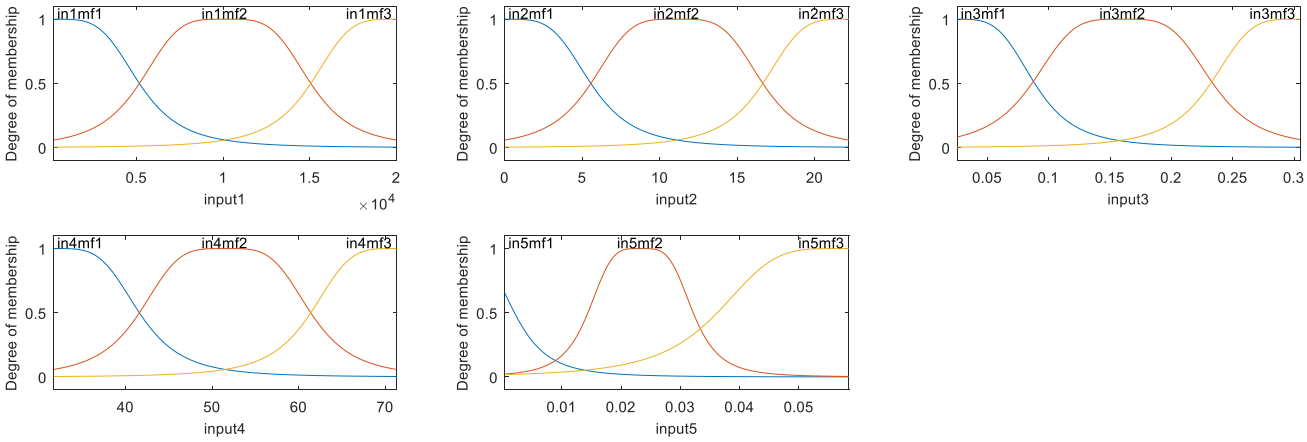
Ζητούμενα του προβλήματος:

1. Παρατίθενται τα διαγράμματα στα οποία απεικονίζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν μέσω της διαδικασίας εκπαίδευσης.

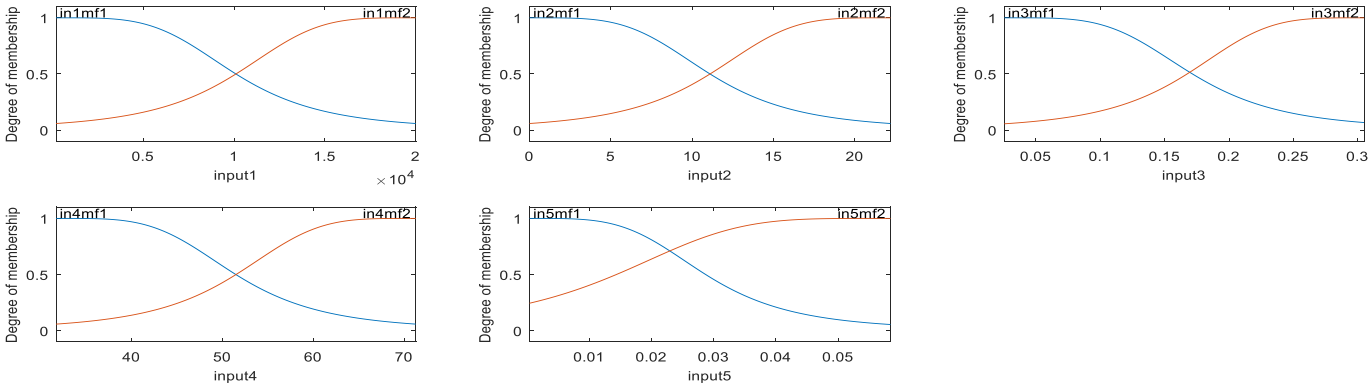
TSK Model 1 (2 membership functions, Singleton output)



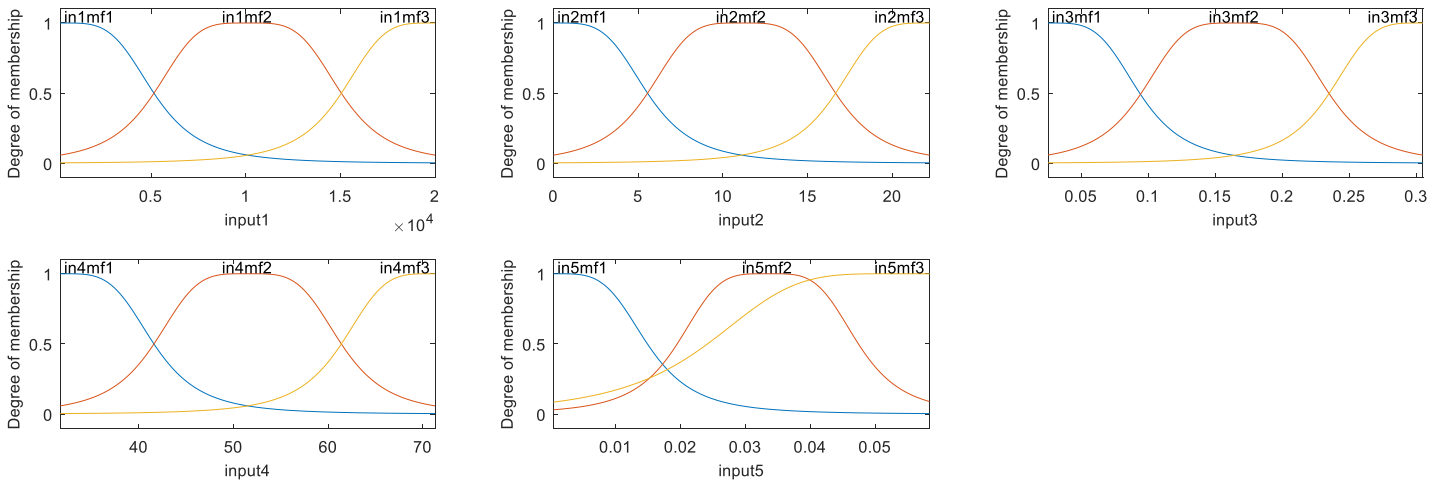
TSK Model 2 (3 membership functions, Singleton output)



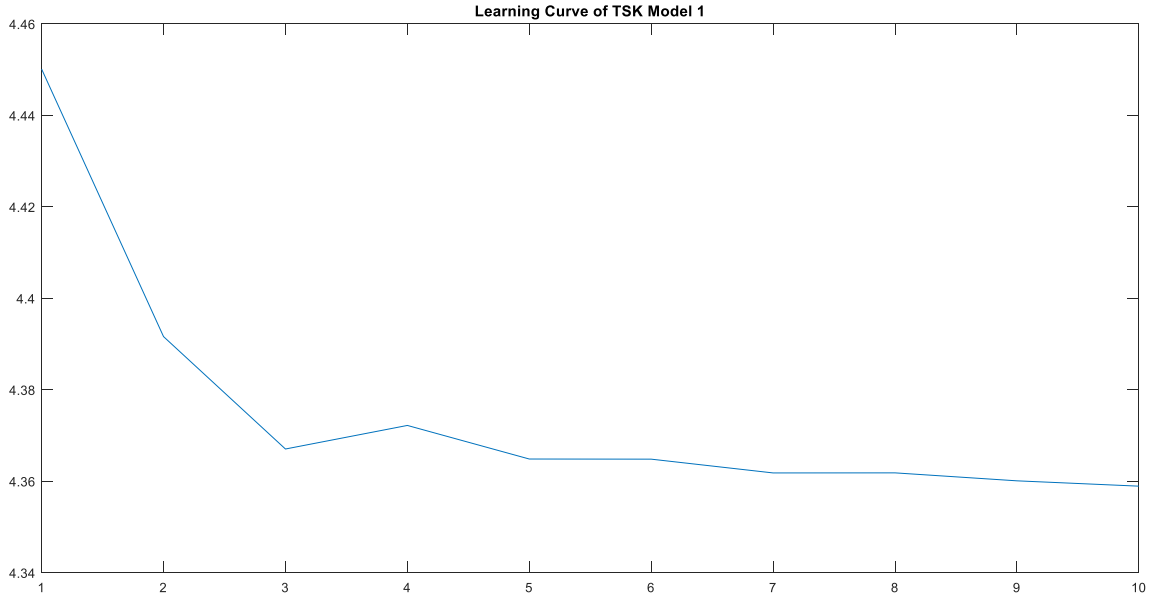
TSK Model 3 (2 membership functions, Polynomial output)

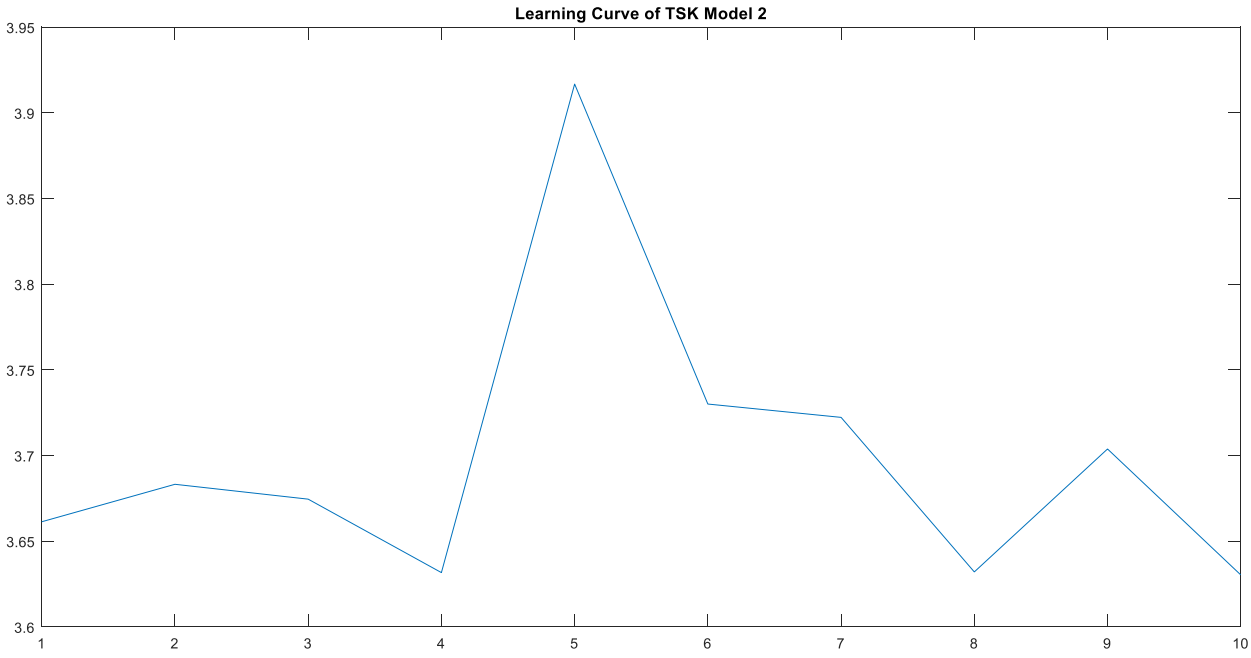


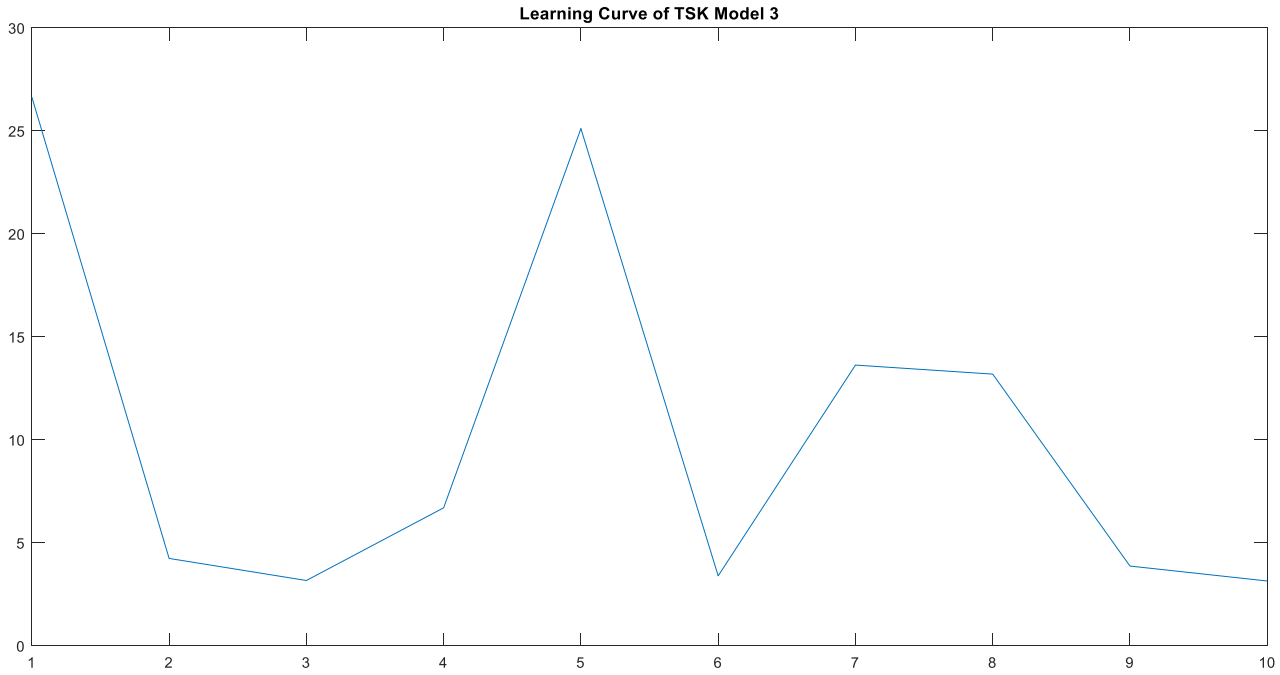
TSK Model 4 (3 membership functions, Polynomial output)

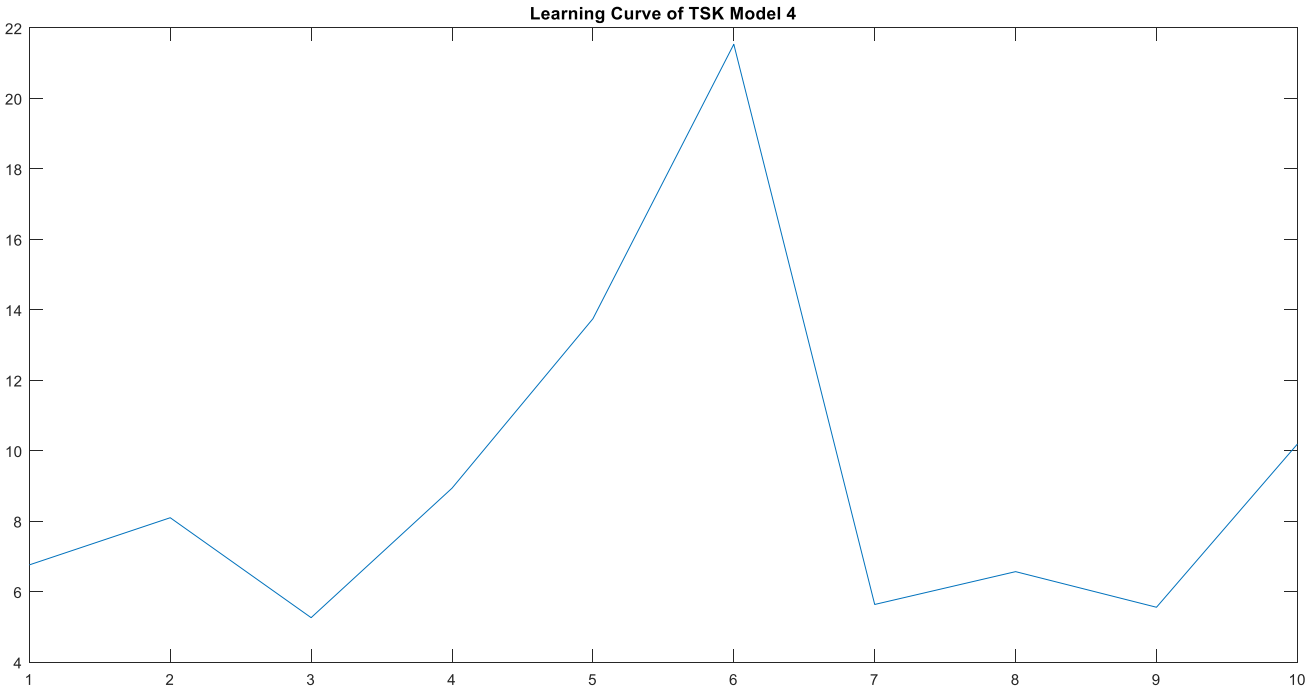


2. Παρατίθενται τα διαγράμματα μάθησης (learning curves) όπου απεικονίζεται το σφάλμα του μοντέλου συναρτήσει του αριθμού των επαναλήψεων (iterations).

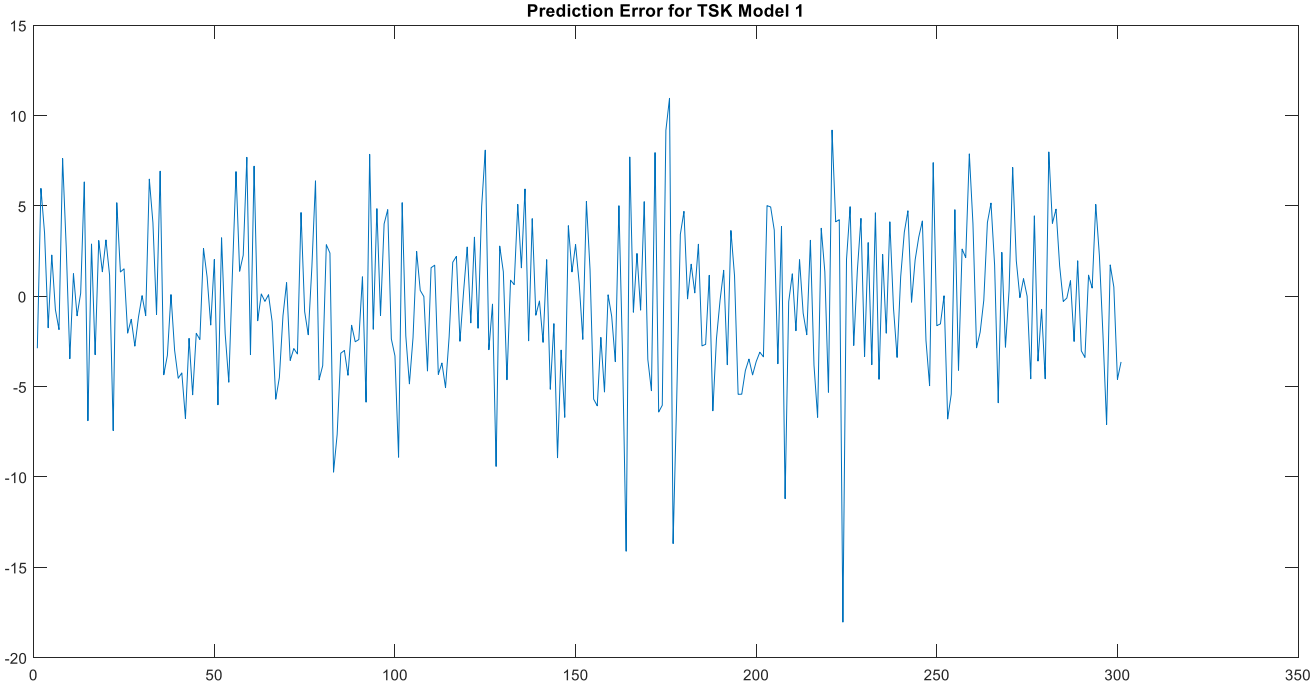


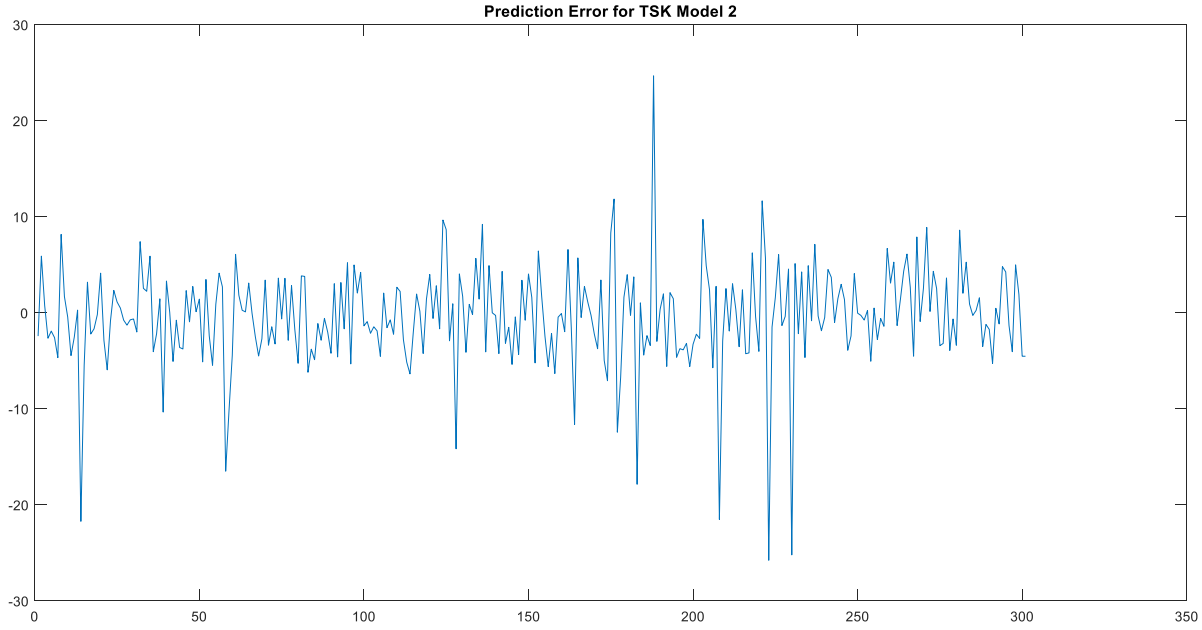


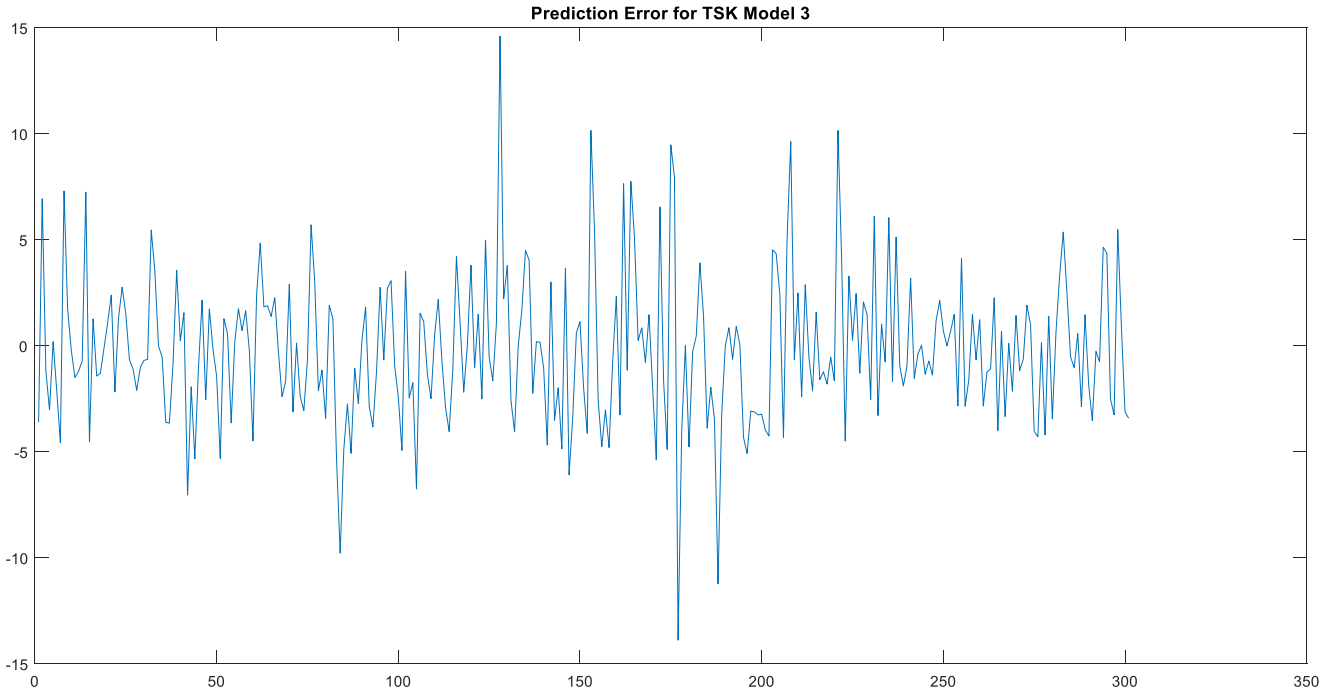

Learning Curve of TSK Model 3
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10

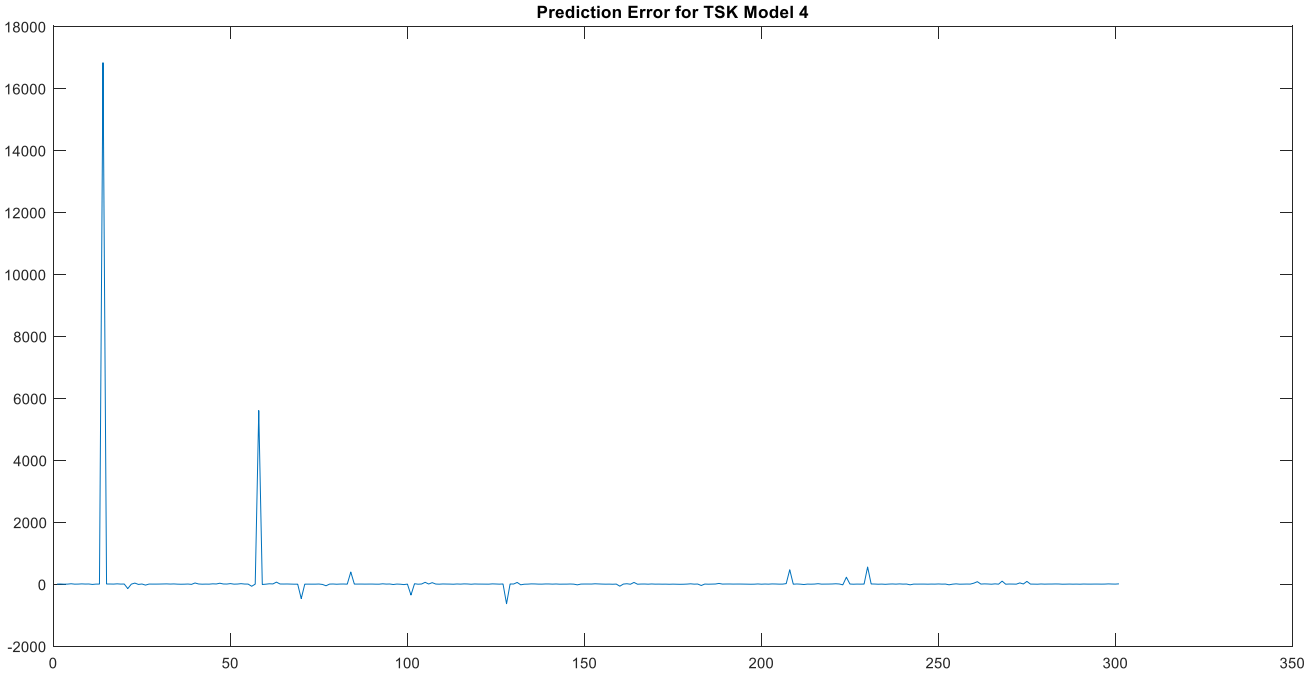

Learning Curve of TSK Model 4
22
20
18
16
14
12
10
8
6
4
1
2
3
4
5
6
7
8
9
10

3. Δίνονται τα διαγράμματα όπου αποτυπώνονται τα σφάλματα πρόβλεψης.






Prediction Error for TSK Model 3
15
10
5
0
-5
-10
-15
0
50
100
150
200
250
300
350


Prediction Error for TSK Model 4
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
-2000
0
50
100
150
200
250
300
350

4. Παρουσιάζονται σε μορφή πίνακα οι τιμές των δεικτών απόδοσης RMSE, NMSE, NDEI και R^2.

	MSE	RMSE	R2	NMSE	NDEI
TSK_model_1	18.216	4.268	0.61836	0.38164	0.61777
TSK_model_2	27.459	5.2401	0.4247	0.5753	0.75848
TSK_model_3	12.042	3.4702	0.7477	0.2523	0.5023
TSK_model_4	1.0498e+06	1024.6	-21994	21995	148.31

Φαίνεται ότι τους καλύτερους δείκτες απόδοσης τους πετυχαίνουμε για το *3ο TSK μοντέλο*, το οποίο έχει 2 συναρτήσεις συμμετοχής και πολυωνυμική μορφή εξόδου. Ο συνδυασμός αυτός δείχνει να είναι βέλτιστος στην περίπτωσή μας, με ένα μοντέλο που είναι πιο σύνθετο τόσο όσο, ώστε να αποφεύγετε και η μεγάλη πολυπλοκότητα και το overtraining.

Βλέπουμε ότι στο 4ο μοντέλο, η τιμή του R^2 είναι αρνητική. Αυτό συμβαίνει λόγω του dataset, η δομή του οποίου δεν χαρακτηριζόταν από ομοιόμορφη κατανομή των τιμών των χαρακτηριστικών του, παρά τις προσπάθειες που έγιναν για να μπορέσει να γίνει ομοιόμορφη δειγματοληψία. Για το λόγο αυτό εμφανίζεται αρνητική τιμή στη μετρική R^2 .

2 Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Στο δεύτερο μέρος της εργασίας, επιλέγεται από το UCI repository το Superconductivity dataset[2], το οποίο περιλαμβάνει 21263 δείγματα και 81 χαρακτηριστικά.

Λόγω του μεγέθους των διαστάσεων των δειγμάτων είναι φανερό ότι μια απλή εφαρμογή του TSK μοντέλου καθίσταται απαγορευτική, ένα πρόβλημα γνωστό και ως κατάρα των διαστάσεων (curse of dimensionality). Για αυτό το λόγο θα χρησιμοποιήσουμε μεθόδους ομαδοποίησης των δεδομένων και ουσιαστικά μείωσης της διασπασιμότητας, καθώς και του αριθμού των IF-THEN κανόνων. Αυτό θα γίνει μέσω της επιλογής χαρακτηριστικών και της διαμέρισης διασκορπισμού.

Στη συνέχεια ακολουθούμε τα εξής βήματα:

- ✓ *Διαχωρίζουμε τα δεδομένα σε σύνολα εκπαίδευσης-επικύρωσης-ελέγχου*

  Ακολουθούμε την ίδια διαδικασία όπως και στο πρώτο μέρος της εργασίας.

  Ελέγχουμε επίσης αν στο dataset υπάρχουν missing values, επομένως και ανάγκη διαχείρισης αυτών.
  Βλέπουμε στο *Abstract: Missing Values?-N/A* και εξετάζοντας το dataset παρατηρούμε ότι σε ορισμένα χαρακτηριστικά υπάρχει η τιμή 0, αν και δεν έχει κάποια φυσική σημασία (π.χ., standard atomic mass). Υπάρχει ενδεχομένως η πιθανότητα στο dataset οι άγνωστες τιμές να αντιπροσωπεύονται με 0, κάτι το οποίο μπορεί να χειροτερέψει την εκπαίδευση του μοντέλου μας. Βρέθηκαν 215 (215/21263 = 1.01% των δειγμάτων) δείγματα με τουλάχιστον ένα

[2] https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data

μηδενικό σε κάποιο χαρακτηριστικό τους. Αφαιρώντας τα από το dataset, παρατηρήθηκε μια μικρή βελτίωση στο μοντέλο μας.

- ✓ *Επιλέγουμε τις βέλτιστες παραμέτρους*

  Στο σύστημά μας έχουμε δύο ελεύθερες παραμέτρους, τον αριθμό των χαρακτηριστικών και την ακτίνα των clusters r_a. Θα χρησιμοποιήσουμε τη μέθοδο αναζήτησης πλέγματος (grid search) για την εύρεση του βέλτιστου συνδυασμού αυτών. Όπως αναφέρεται και στην εκφώνηση, επιλέγονται ελεύθερα οι τιμές των παραμέτρων που θα εξεταστούν.

  Αρχικά, χρησιμοποιούμε τη συνάρτηση *relieff()* της Matlab, η οποία χρησιμοποιεί τον αλγόριθμο ReliefF για να εντοπίσει τα πιο σημαντικά χαρακτηριστικά των δεδομένων.

  Στη συνέχεια, χρησιμοποιούμε τη *cvpartition()* για να πετύχουμε διαχωρισμό των δεδομένων μέσω 5-πτυχης διασταυρωμένης επικύρωσης (5-fold cross validation), έτσι ώστε σε κάθε επανάληψη, το 80% των δεδομένων να χρησιμοποιείται για εκπαίδευση και το υπόλοιπο 20% για επικύρωση (η διαμέριση αυτή είναι και η default στη συγκεκριμένη συνάρτηση).

  Χρησιμοποιούμε τη συνάρτηση *genfis2()* για τη δημιουργία του μοντέλου, η οποία κάνει χρήση του αλγορίθμου Subtractive Clustering (SC).

  Εκπαιδεύουμε το μοντέλο χρησιμοποιώντας την *anfis().*

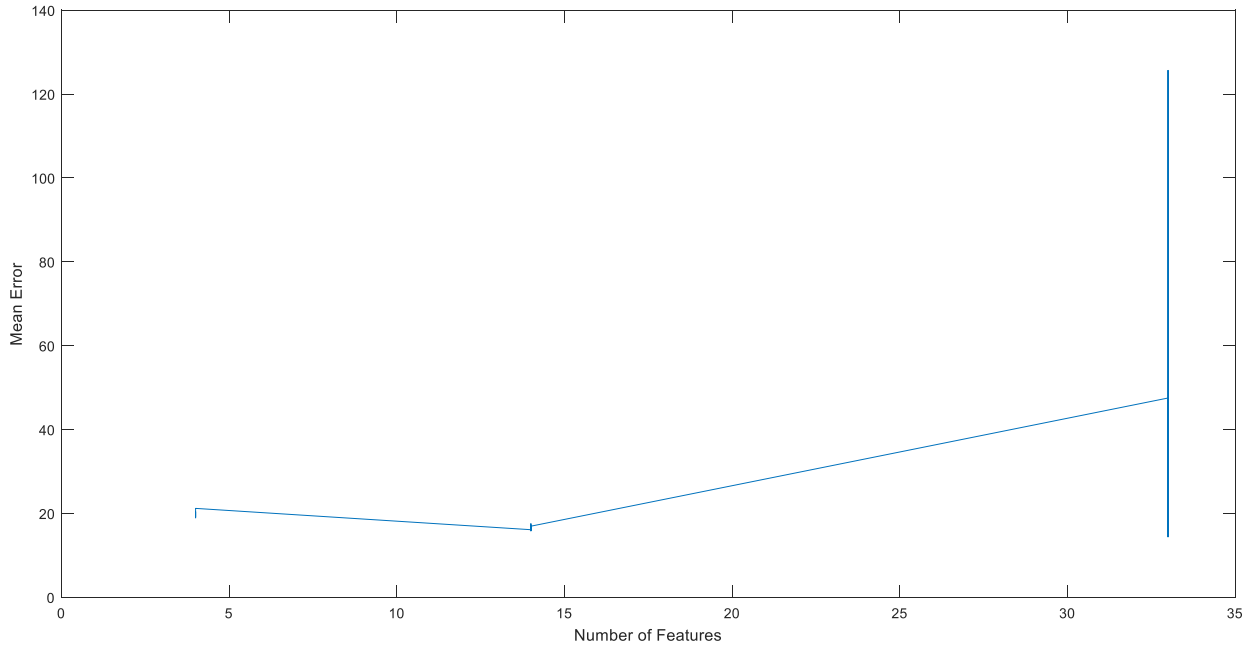
  Αποθηκεύουμε το μέσο σφάλμα σε κάθε μοντέλο που εκπαιδεύουμε και βρίσκουμε τον τελικό βέλτιστο συνδυασμό των ελεύθερων παραμέτρων από το αντίστοιχο ελάχιστο μέσο σφάλμα.

- ✓ *Εκπαιδεύουμε το τελικό TSK μοντέλο και ελέγχουμε την απόδοσή του στο σύνολο ελέγχου*

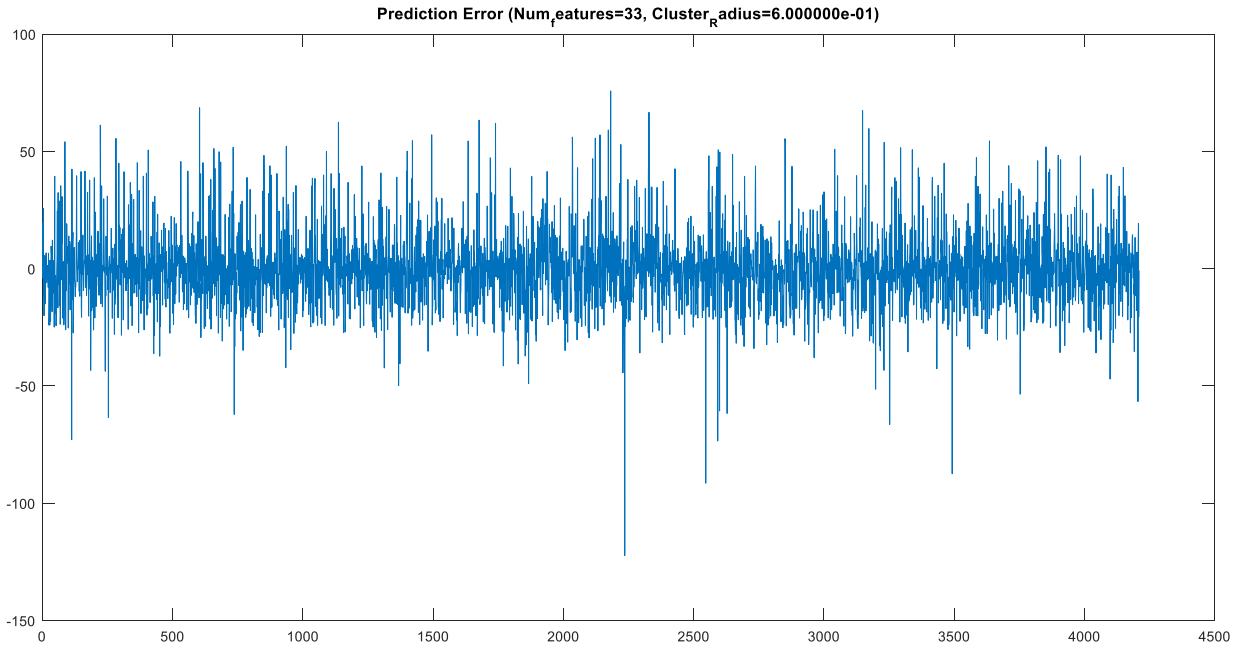
  Χρησιμοποιούμε τη συνάρτηση *genfis2()* για τη δημιουργία του μοντέλου, με τη χρήση πλέον των βέλτιστων χαρακτηριστικών και εκπαιδεύουμε το μοντέλο χρησιμοποιώντας και πάλι την *anfis().*
  Οι βέλτιστες τιμές των παραμέτρων βγήκαν ως *Number of Features = 33* και *Cluster Radius = 0.6*.

Ζητούμενα του προβλήματος:

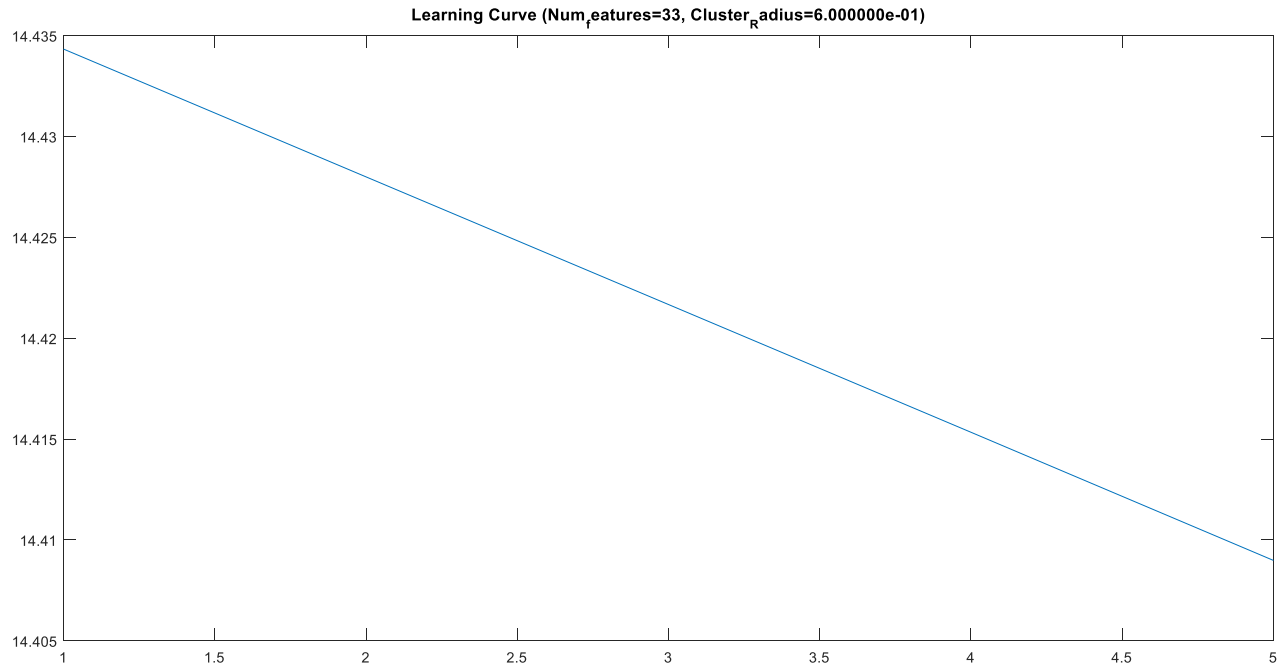
1. Δίνονται το διάγραμμα το οποίο απεικονίζει την καμπύλη του μέσου σφάλματος σε σχέση με τον αριθμό των επιλεχθέντων χαρακτηριστικών.



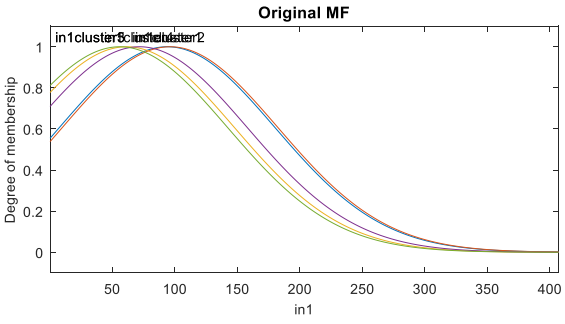
2. Παρατίθεται το διάγραμμα του Prediction Error για τις βέλτιστες παραμέτρους.

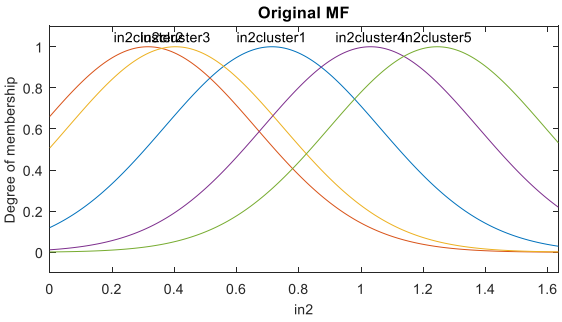


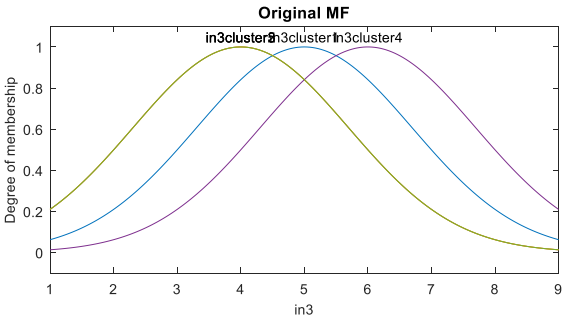
3. Παρουσιάζεται το διάγραμμα εκμάθησης όπου απεικονίζεται το σφάλμα συναρτήσει του αριθμού επαναλήψεων.

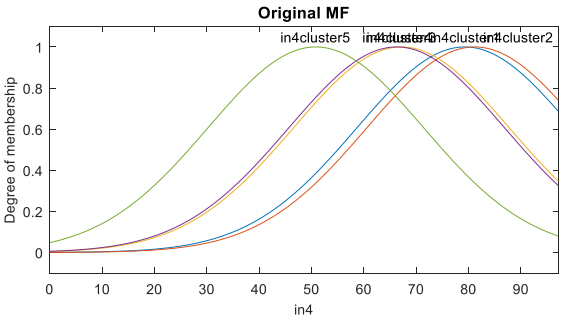


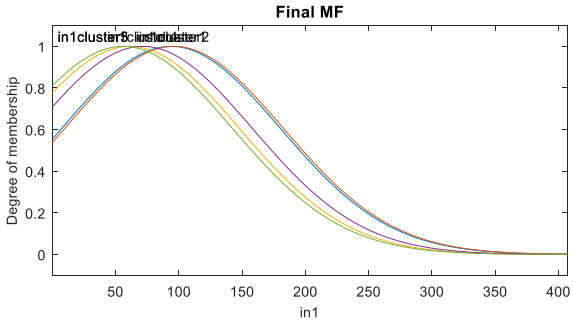
4. Δίνονται ενδεικτικά μερικά ασαφή σύνολα στην αρχική και τελική τους μορφή.

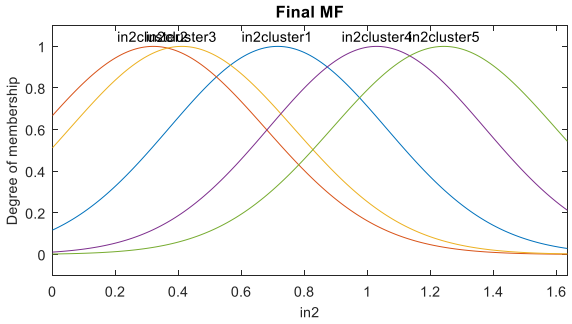


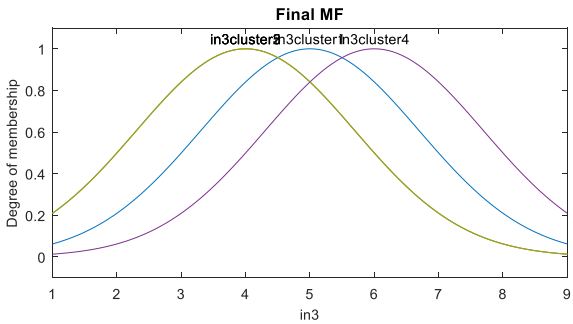


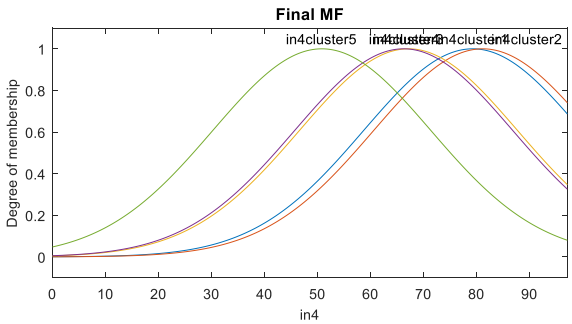












5. Παρουσιάζονται σε μορφή πίνακα οι τιμές των δεικτών απόδοσης RMSE, NMSE, NDEI και R^2.

MSE	RMSE	R2	NMSE	NDEI
227.5	15.083	0.8068	0.1932	0.43955

Αν είχαμε επιλέξει grid partitioning με δύο ή τρία ασαφή σύνολα στην είσοδο θα είχαμε 2^{81} και 3^{81} κανόνες αντίστοιχα. Πρόκειται λοιπόν για μία καθόλου πρακτική πολυπλοκότητα.