

Υπολογιστική Νοημοσύνη

Αναφορά στα πλαίσια της 3ης εργασίας

Regression



Τσαλαγεώργος Βασίλειος

A.E.M. 8253

Εαρινό εξάμηνο 2021

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

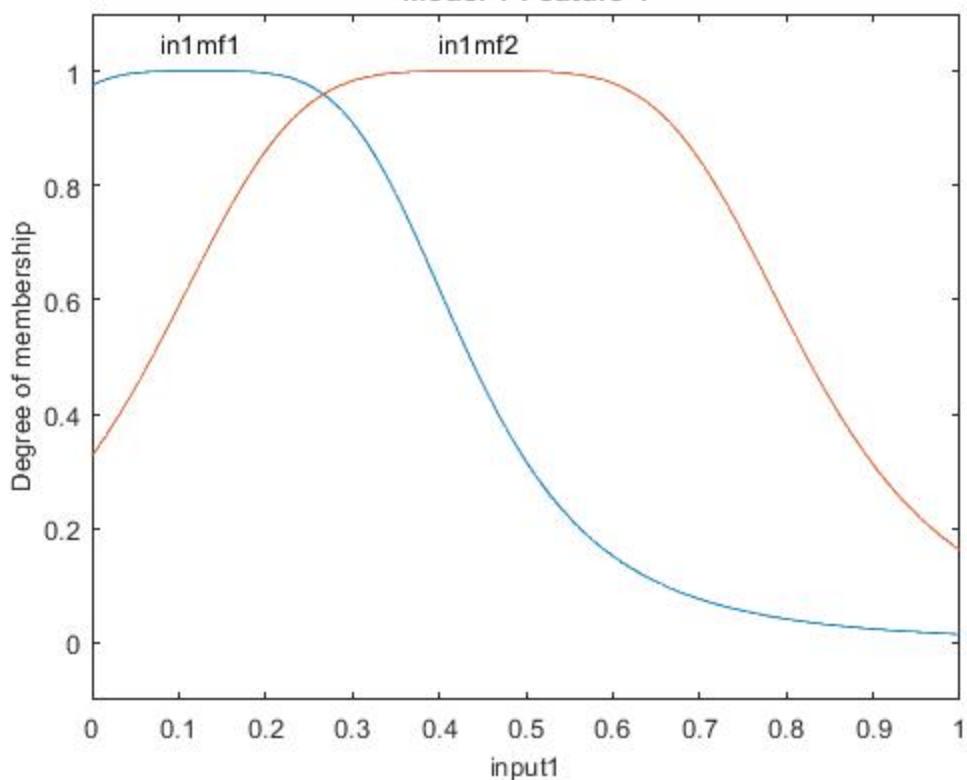
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Εφαρμογή σε απλό dataset

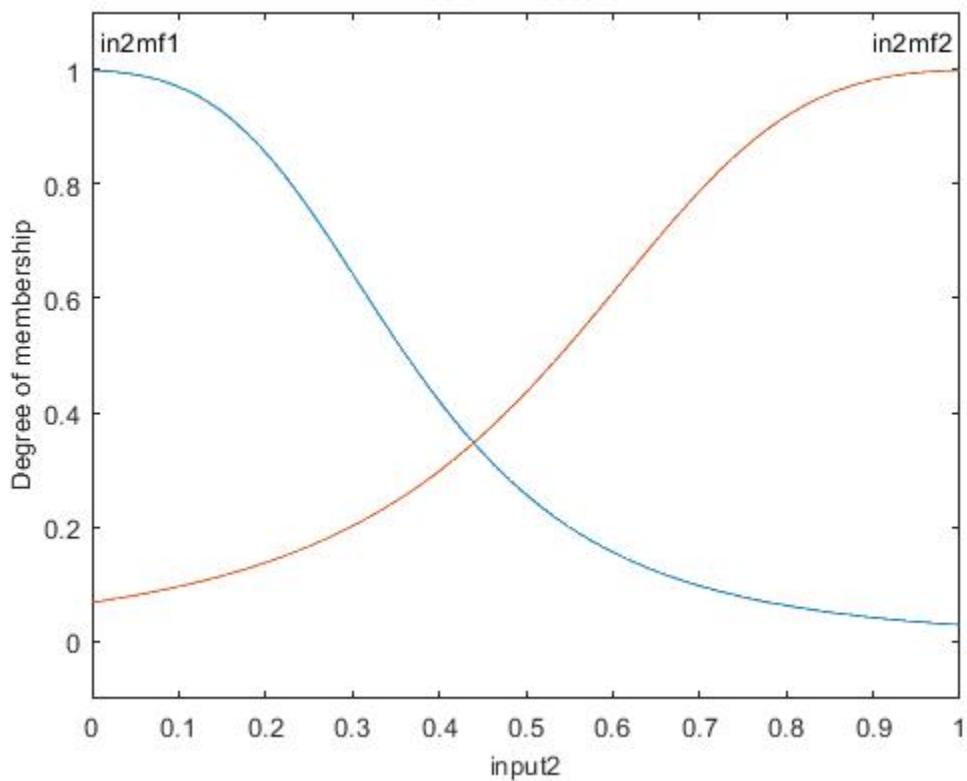
Το πρώτο μέρος της εργασίας υλοποιείται σε κώδικα MATLAB και αποθηκεύεται στο αρχείο main.m. Σε πρώτη φάση, γίνεται ο διαχωρισμός των δεδομένων σε τρία υποσύνολα: εκπαίδευσης (trnData), χρησιμοποιώντας το 60% των συνολικών δεδομένων, επικύρωσης (chkData), χρησιμοποιώντας το 20% των συνολικών δεδομένων, και ελέγχου (tstData), χρησιμοποιώντας το εναπομείναν 20% των συνολικών δεδομένων. Ο διαχωρισμός επιτυγχάνεται με τη χρήση της συνάρτησης που βρίσκεται στο αρχείο split_scale.m. Εν συνεχείᾳ, ορίζονται τα τέσσερα TSK μοντέλα τα οποία πρόκειται να εκπαιδεύσουμε, με τον αριθμό των συναρτήσεων συμμετοχής και την έξοδο του καθενός να ορίζονται σύμφωνα με τον πίνακα 1 της εκφώνησης. Έπειτα, ξεκινά η εκπαίδευση του κάθε μοντέλου για 100 epochs, και σχεδιάζονται οι συναρτήσεις συμμετοχής του καθενός, για κάθε ένα χαρακτηριστικό (feature) των δεδομένων που χρησιμοποιούμε (5 στο σύνολο, το έκτο είναι η έξοδος). Επίσης σχεδιάζονται τα διαγράμματα μάθησης του κάθε μοντέλου (συναρτήσει του αριθμού των επαναλήψεων, όπως ζητείται), καθώς και τα διαγράμματα όπου αποτυπώνονται τα σφάλματα πρόβλεψης (error prediction). Τέλος, υπολογίζονται για κάθε μοντέλο οι τέσσερις ζητούμενοι δείκτες απόδοσης που χρησιμοποιούνται για την αξιολόγηση του κάθε μοντέλου, και αποθηκεύονται σε έναν 4x4 πίνακα, καθώς έχω τέσσερα μοντέλα και τέσσερις δείκτες απόδοσης για το καθένα.

Παρατίθενται παρακάτω τα διαγράμματα στα οποία απεικονίζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν για κάθε μοντέλο και για κάθε χαρακτηριστικό (feature) μέσω της διαδικασίας εκπαίδευσης.

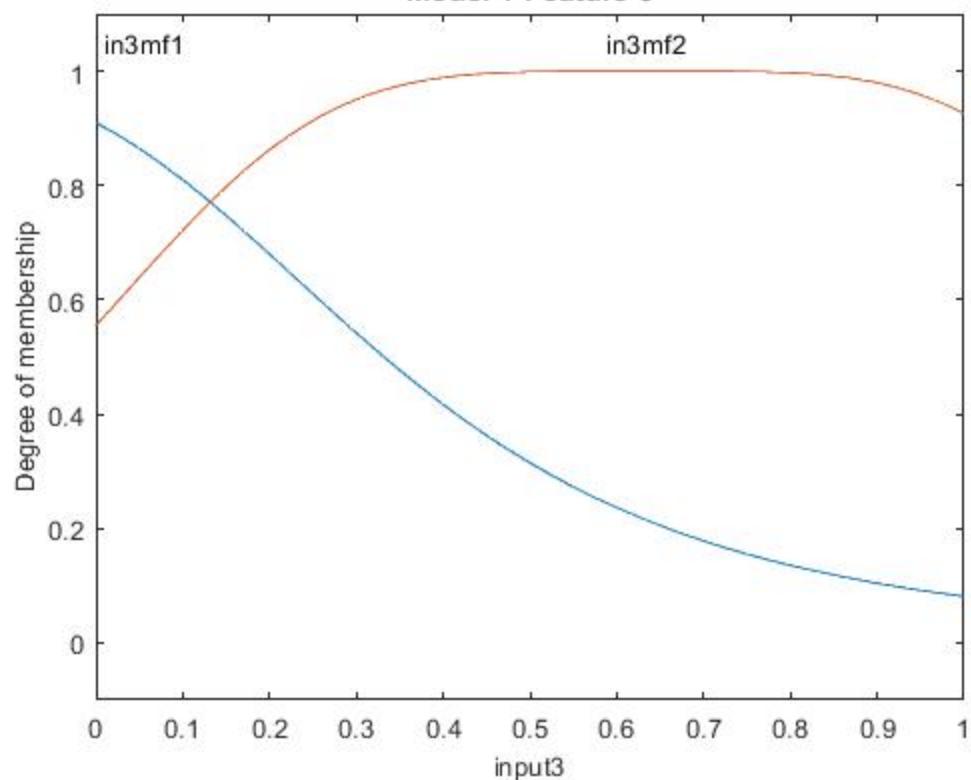
Model 1 Feature 1



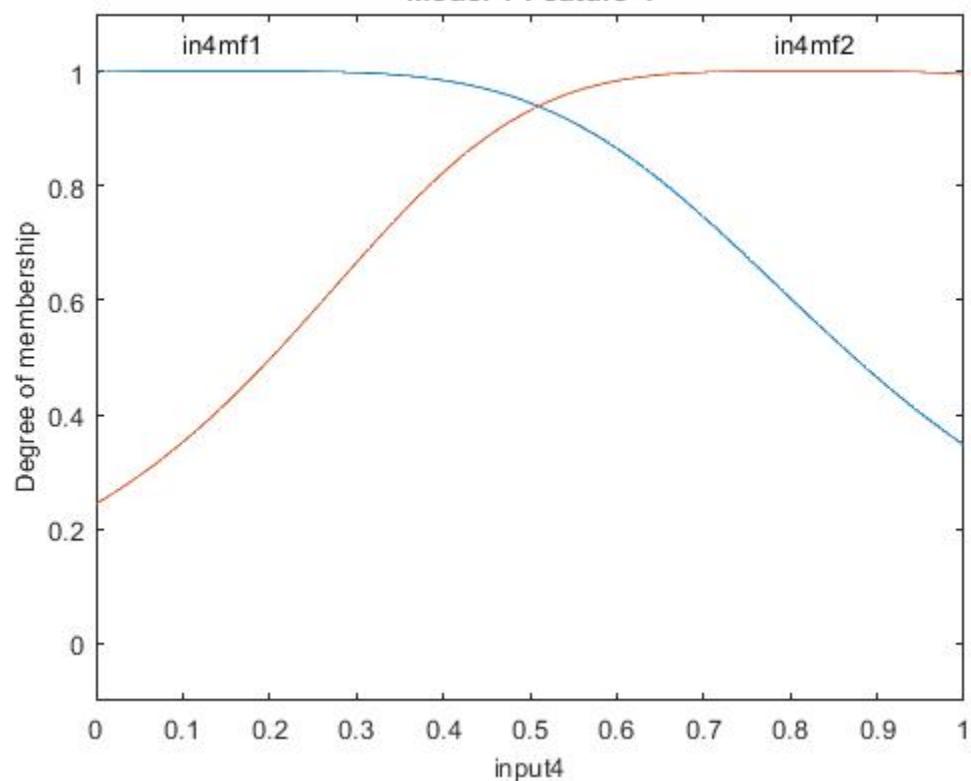
Model 1 Feature 2



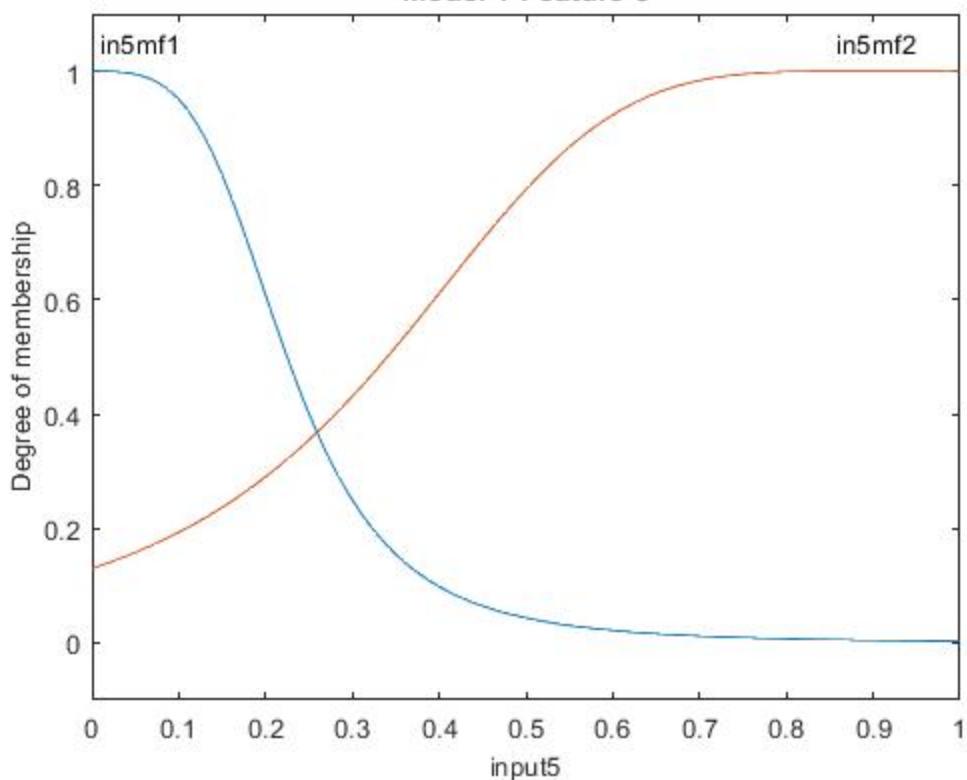
Model 1 Feature 3



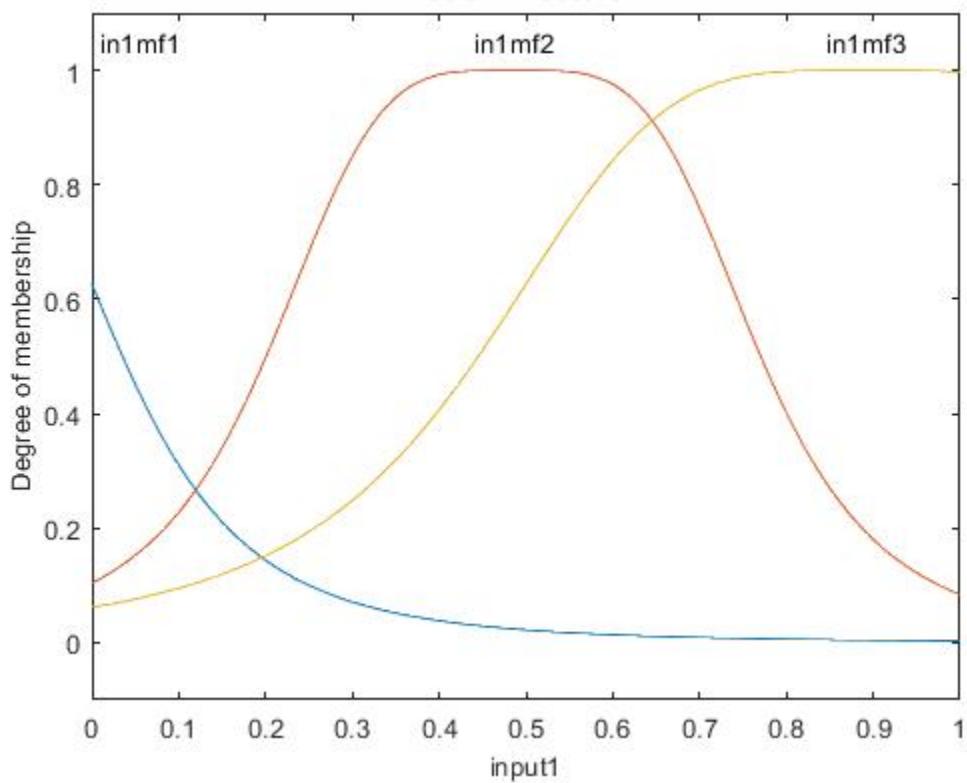
Model 1 Feature 4



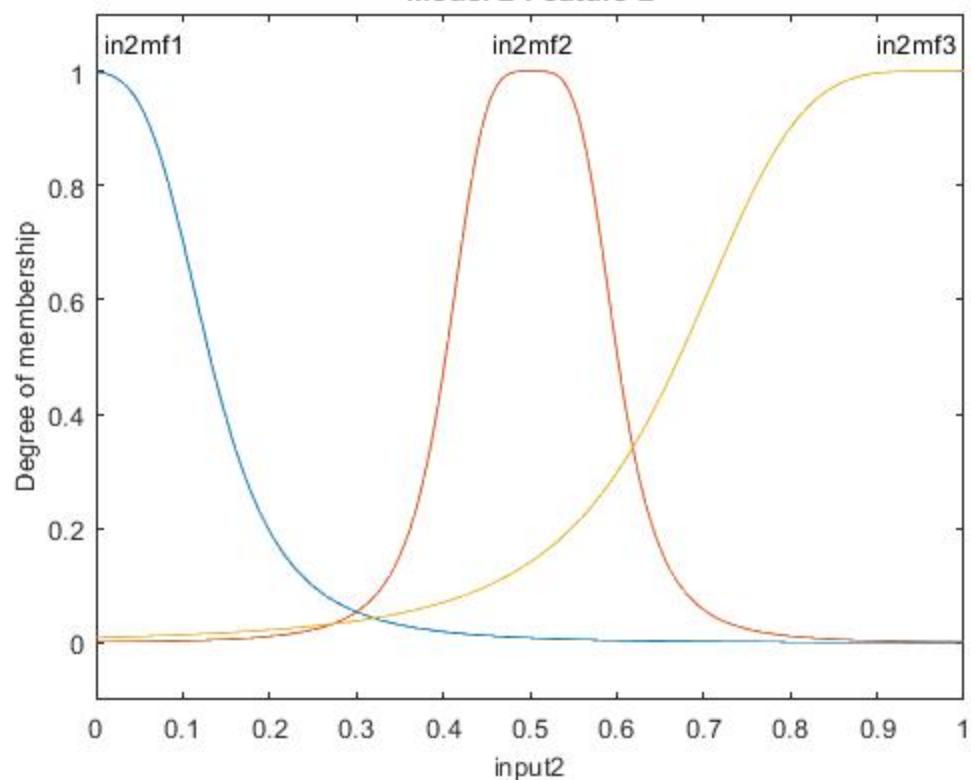
Model 1 Feature 5



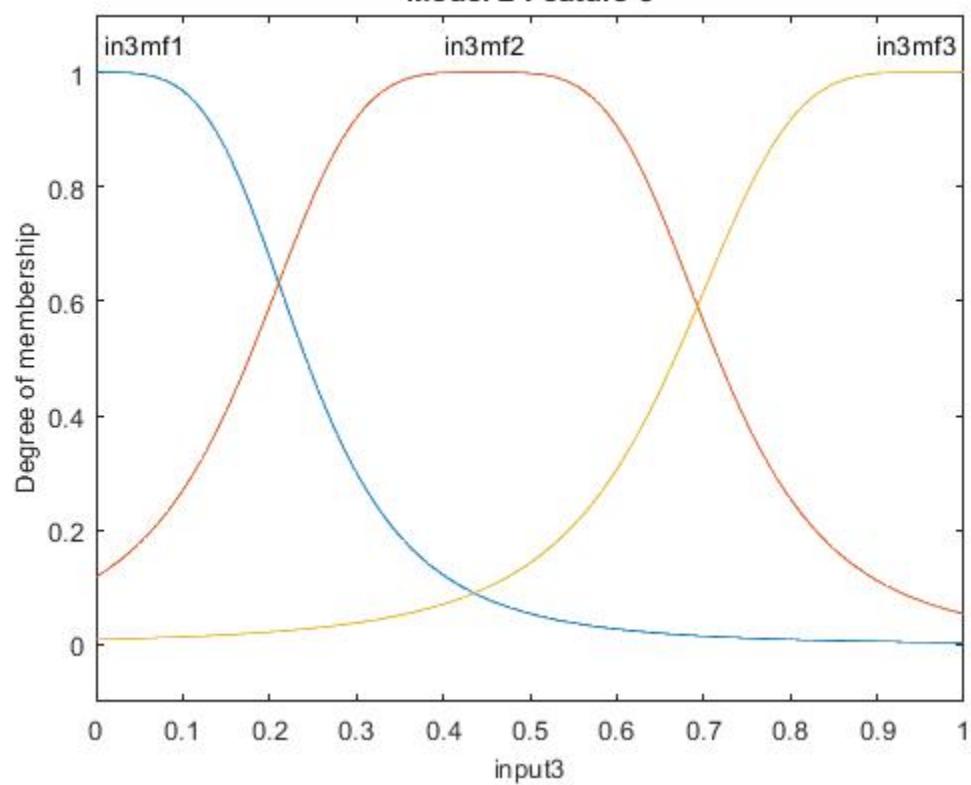
Model 2 Feature 1



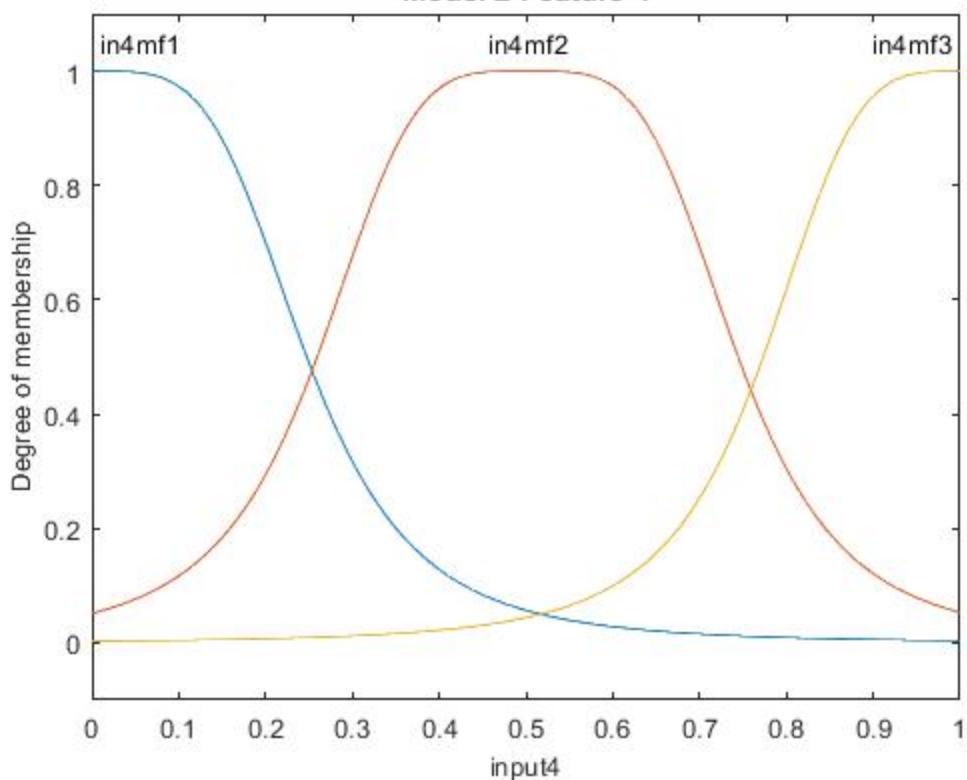
Model 2 Feature 2



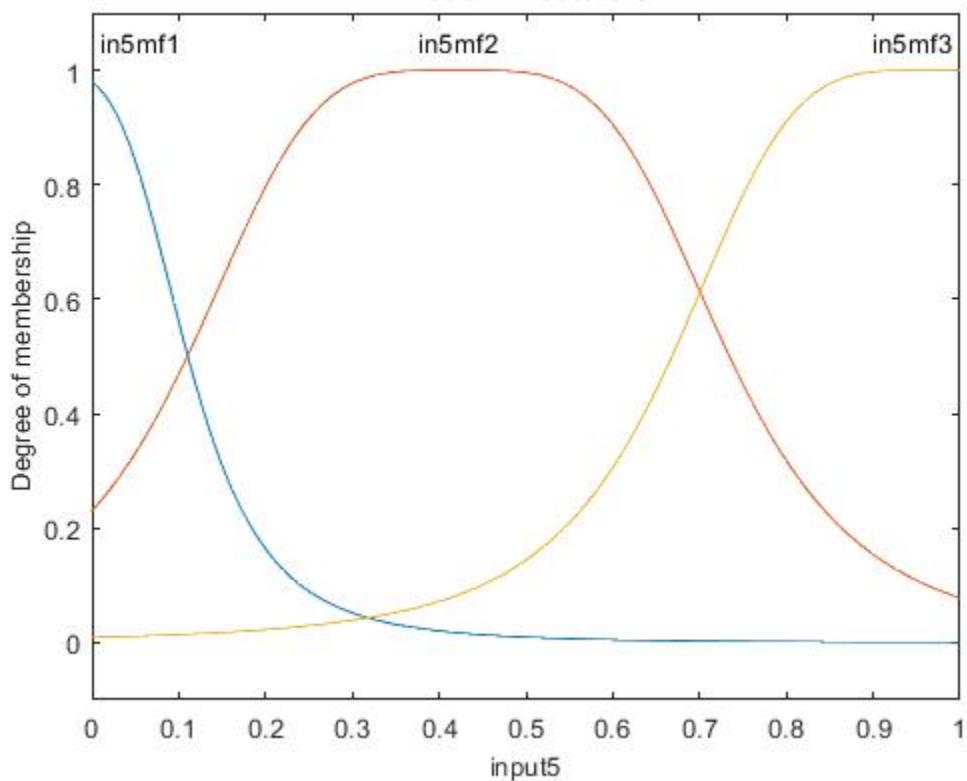
Model 2 Feature 3



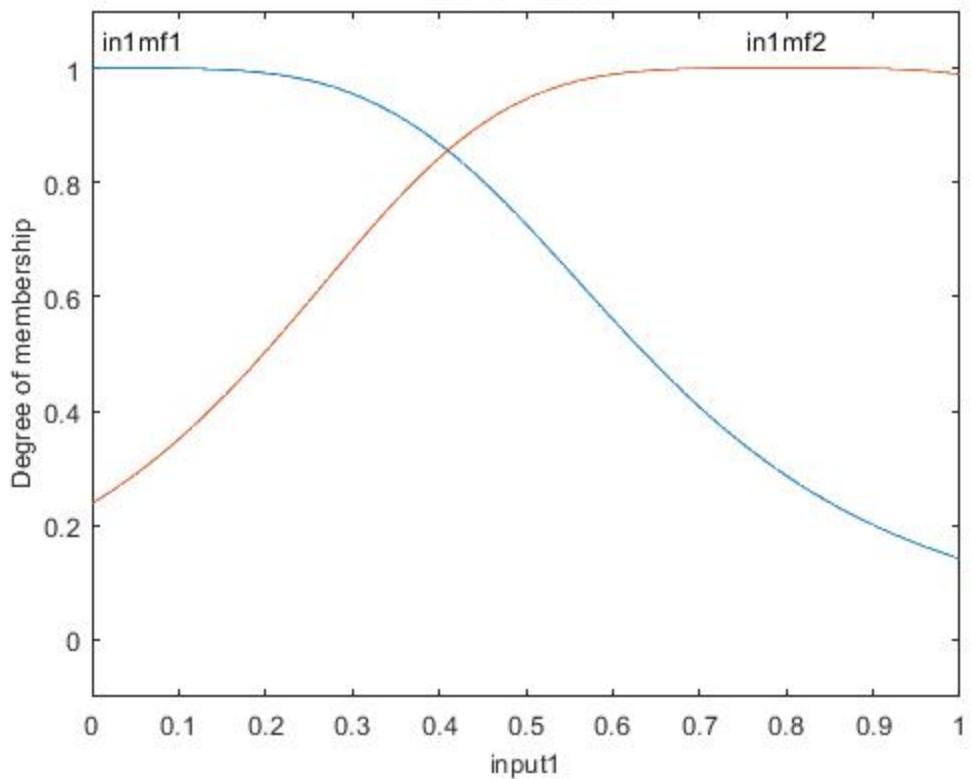
Model 2 Feature 4



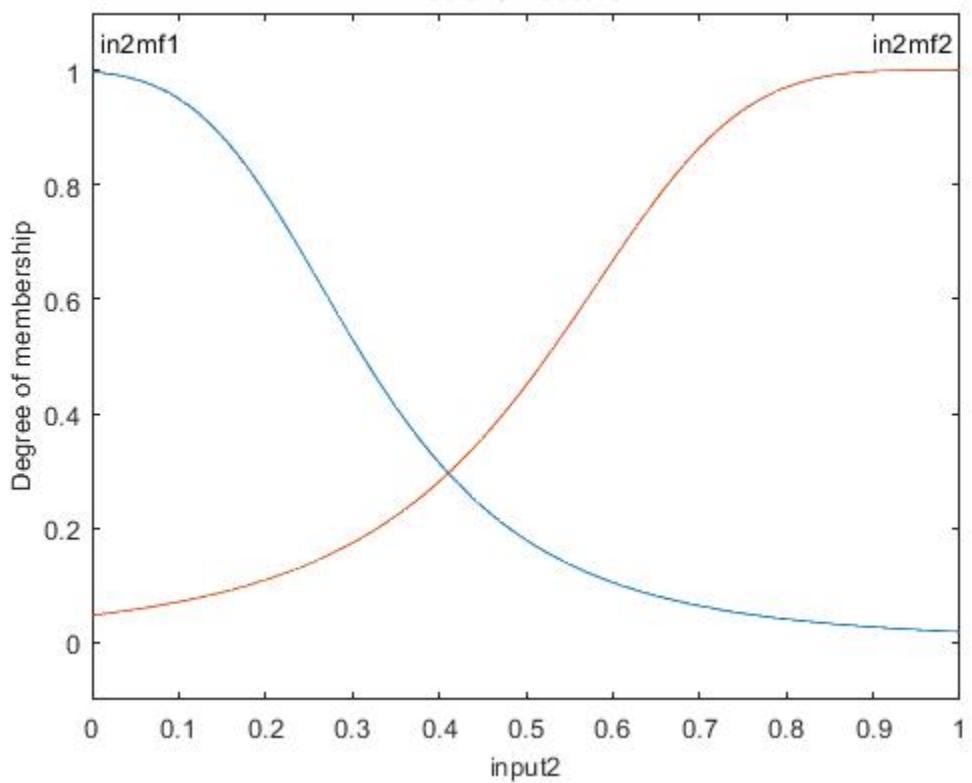
Model 2 Feature 5



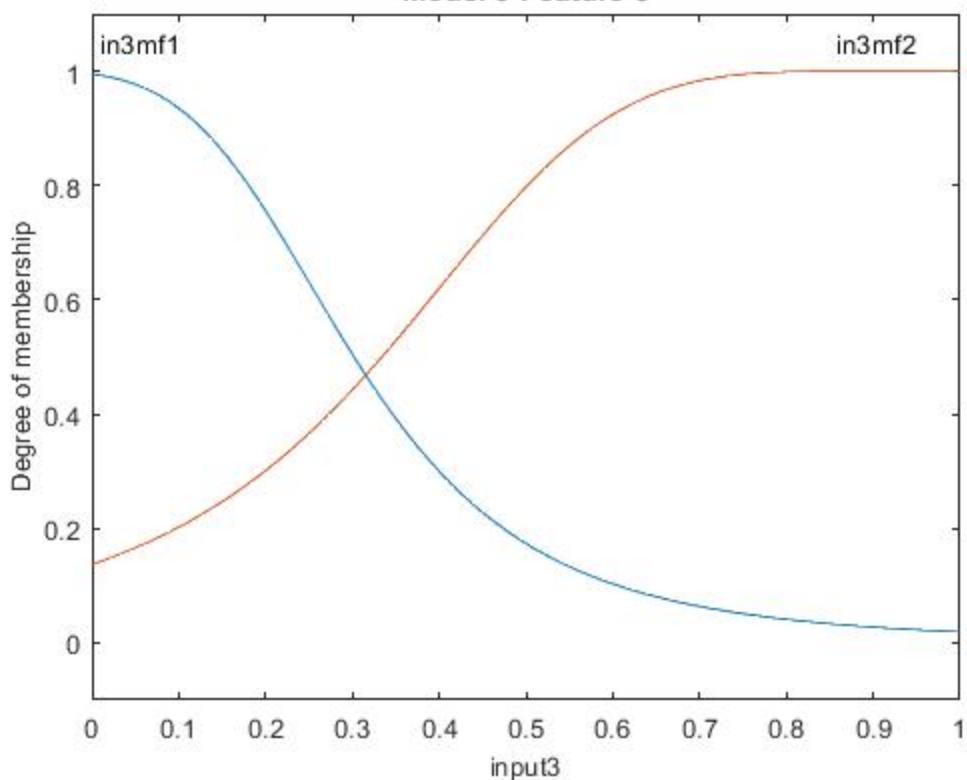
Model 3 Feature 1



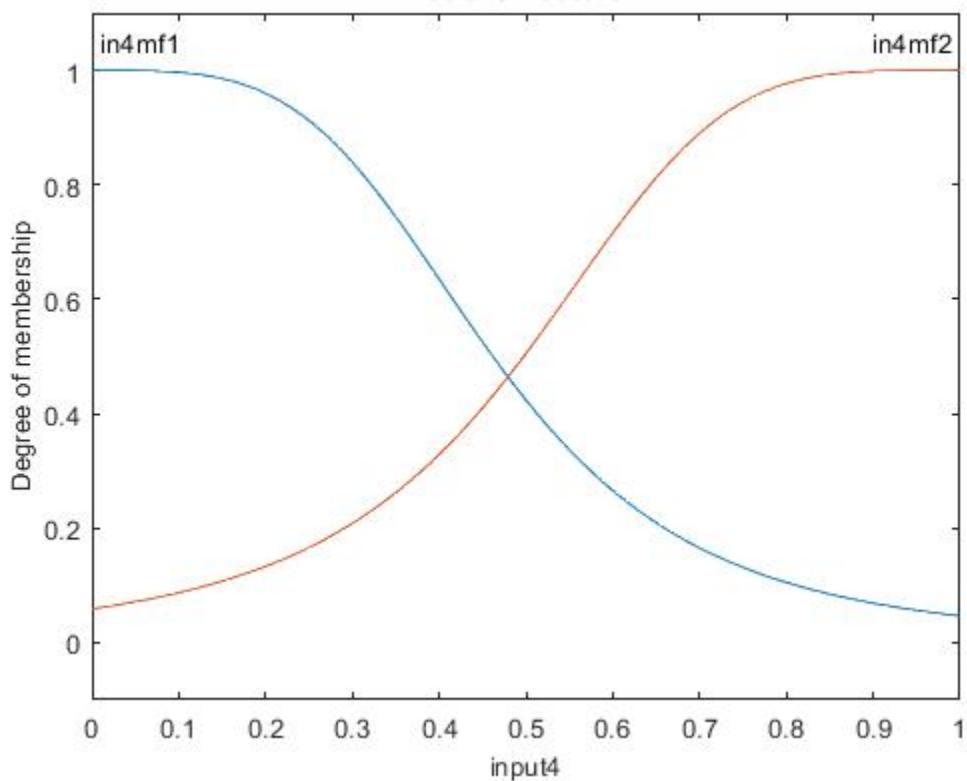
Model 3 Feature 2



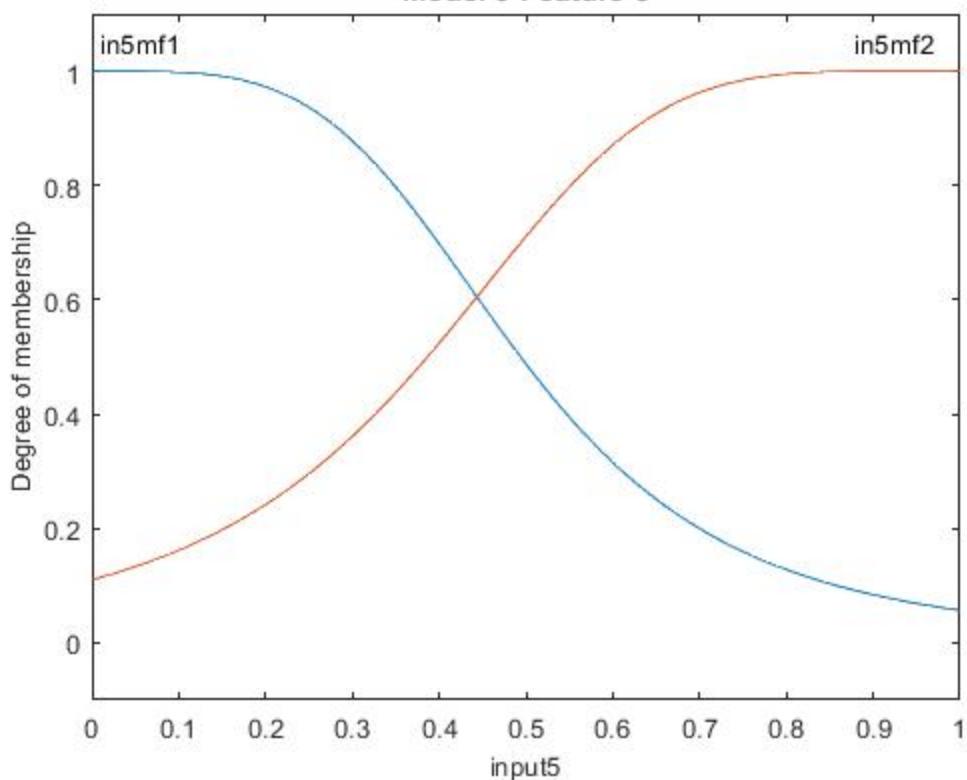
Model 3 Feature 3



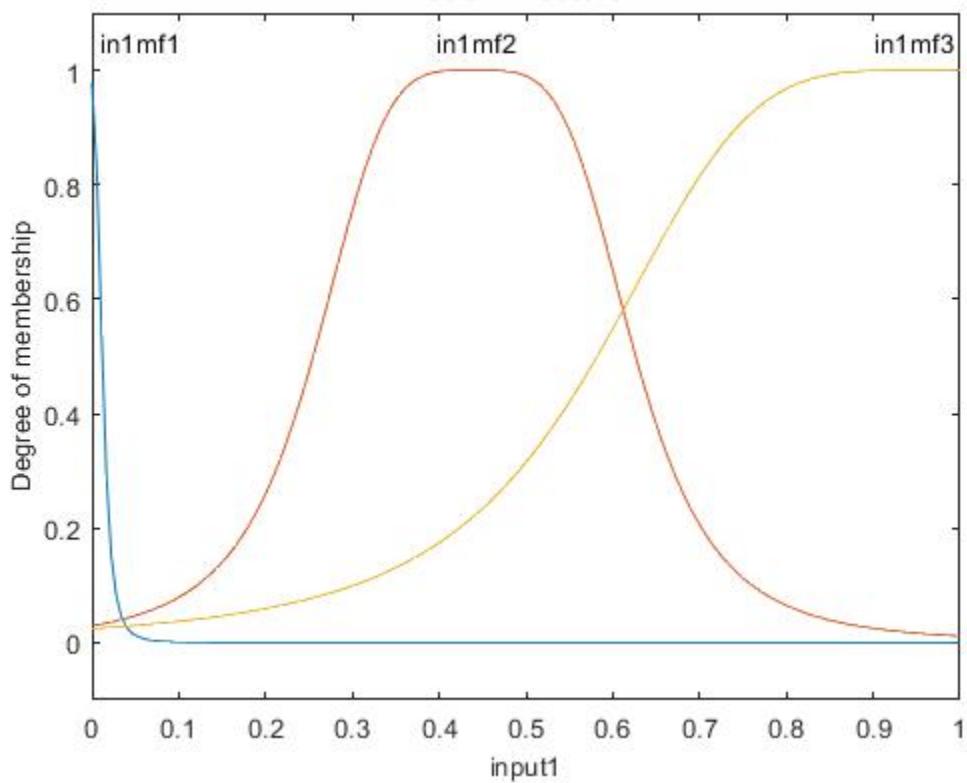
Model 3 Feature 4



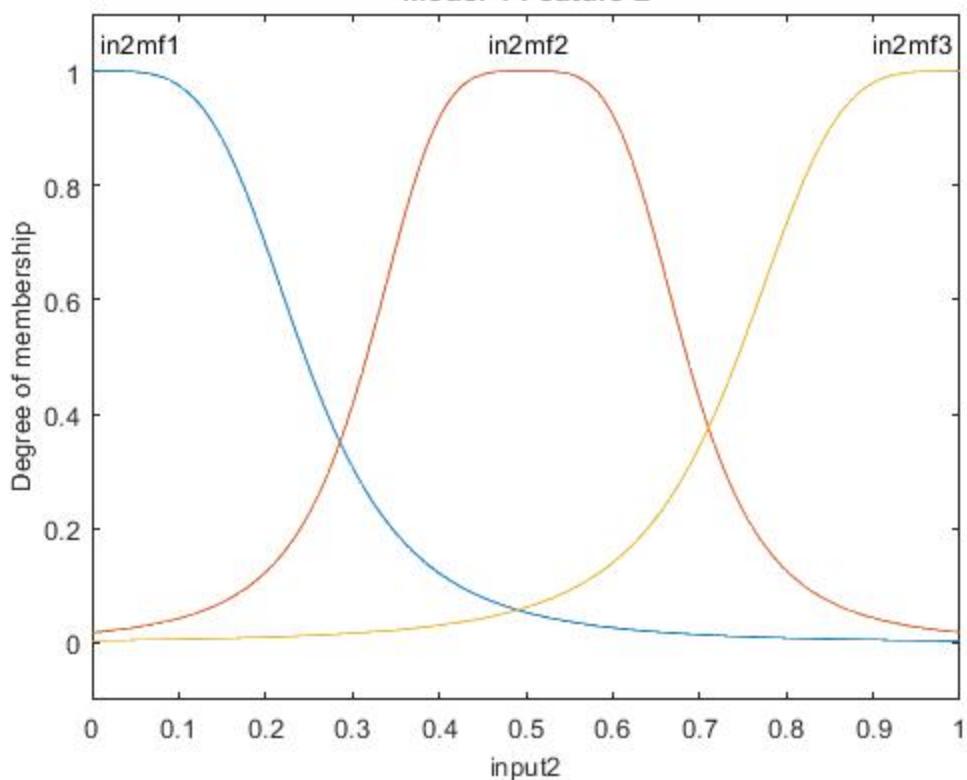
Model 3 Feature 5



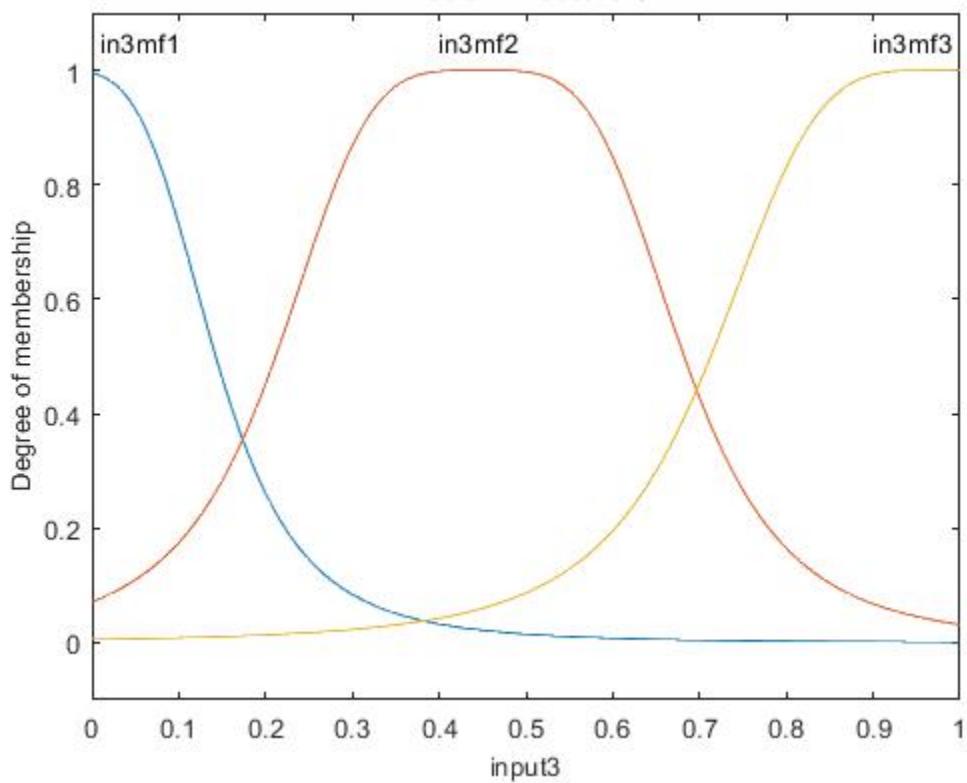
Model 4 Feature 1



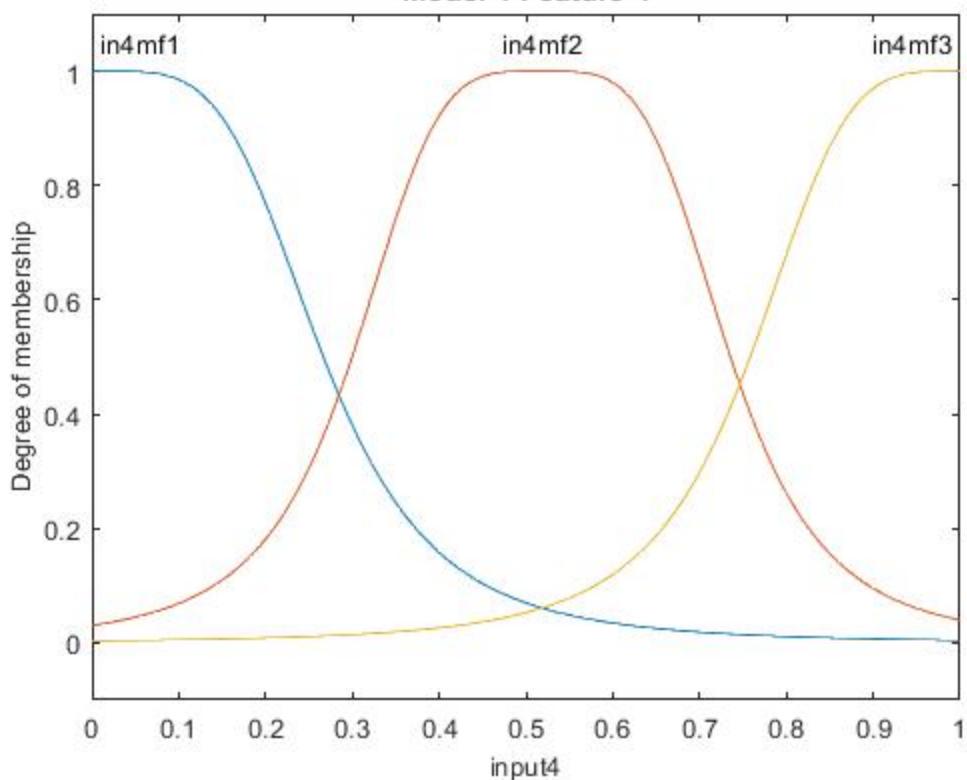
Model 4 Feature 2



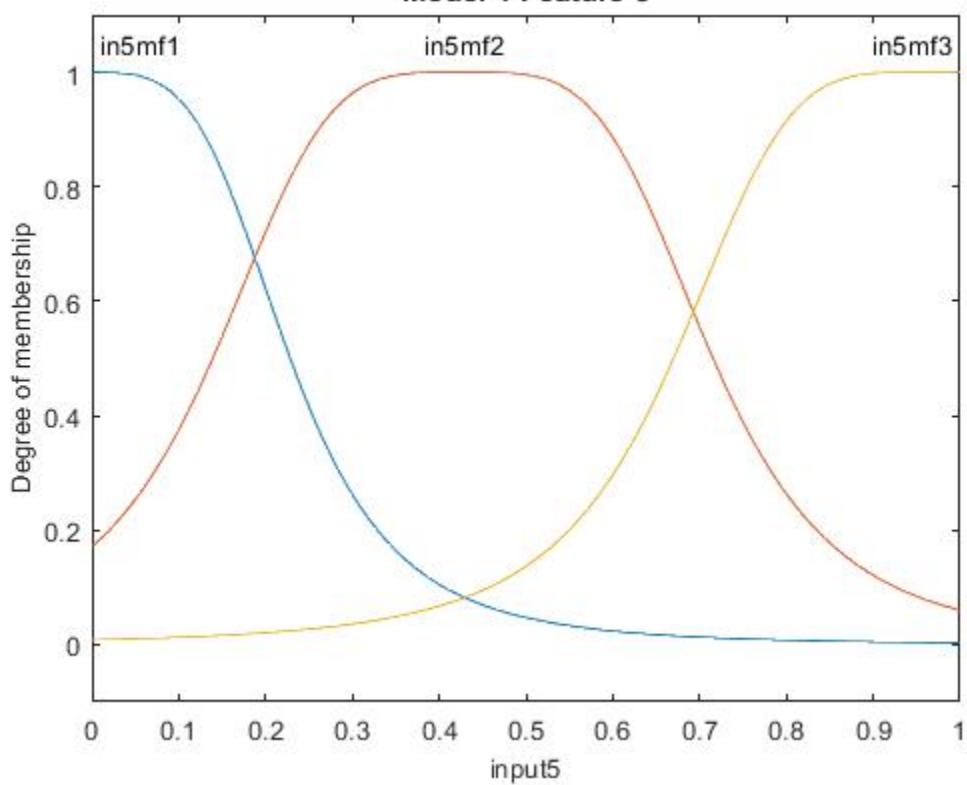
Model 4 Feature 3



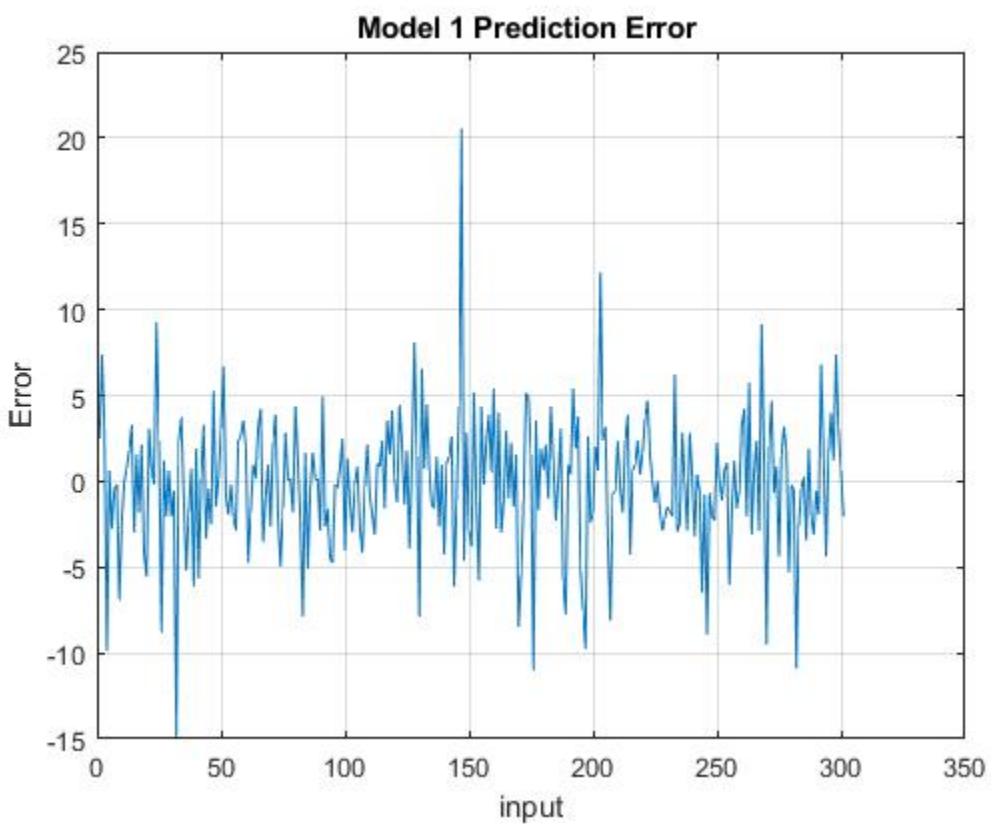
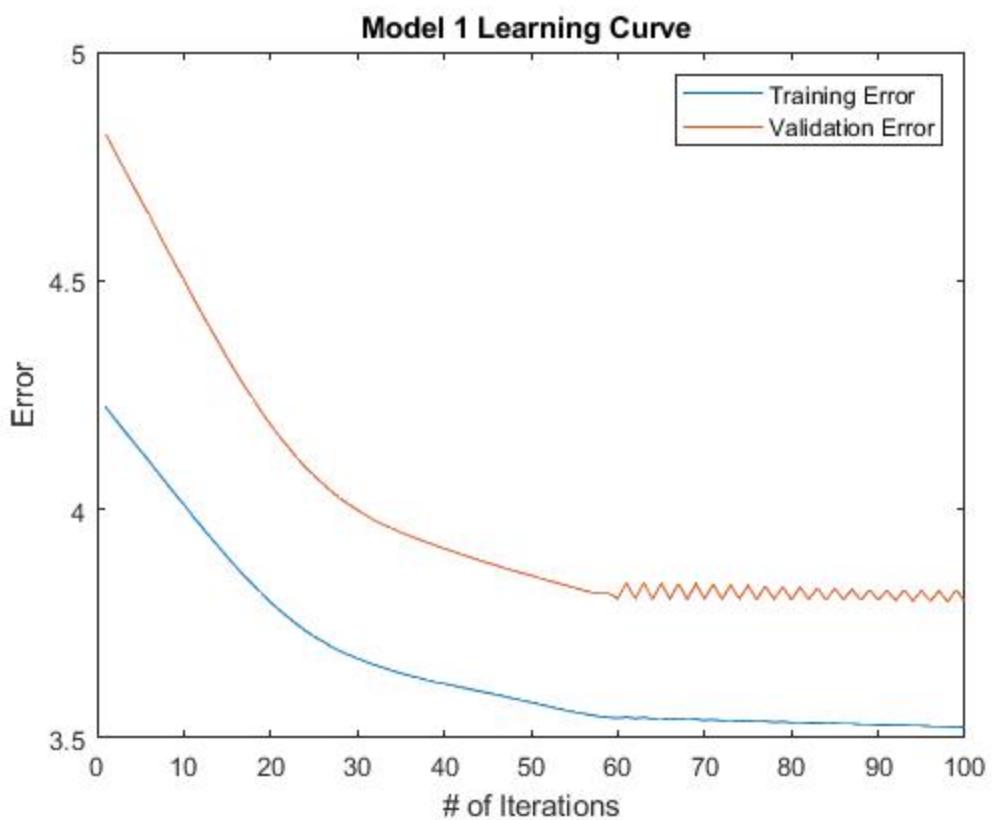
Model 4 Feature 4



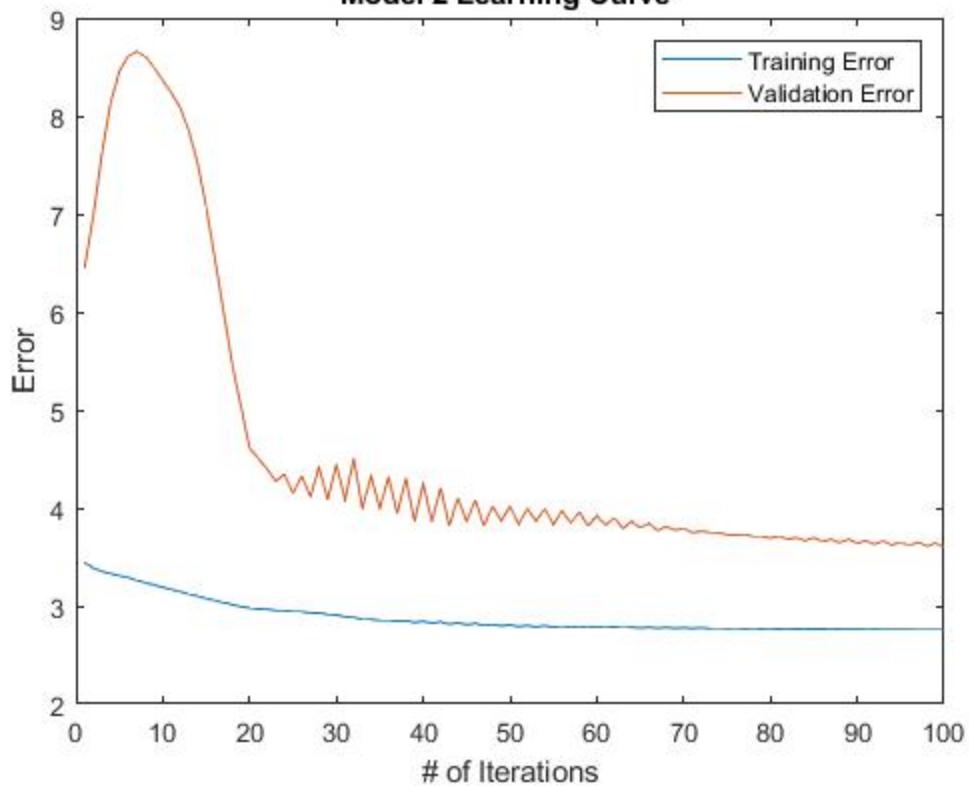
Model 4 Feature 5



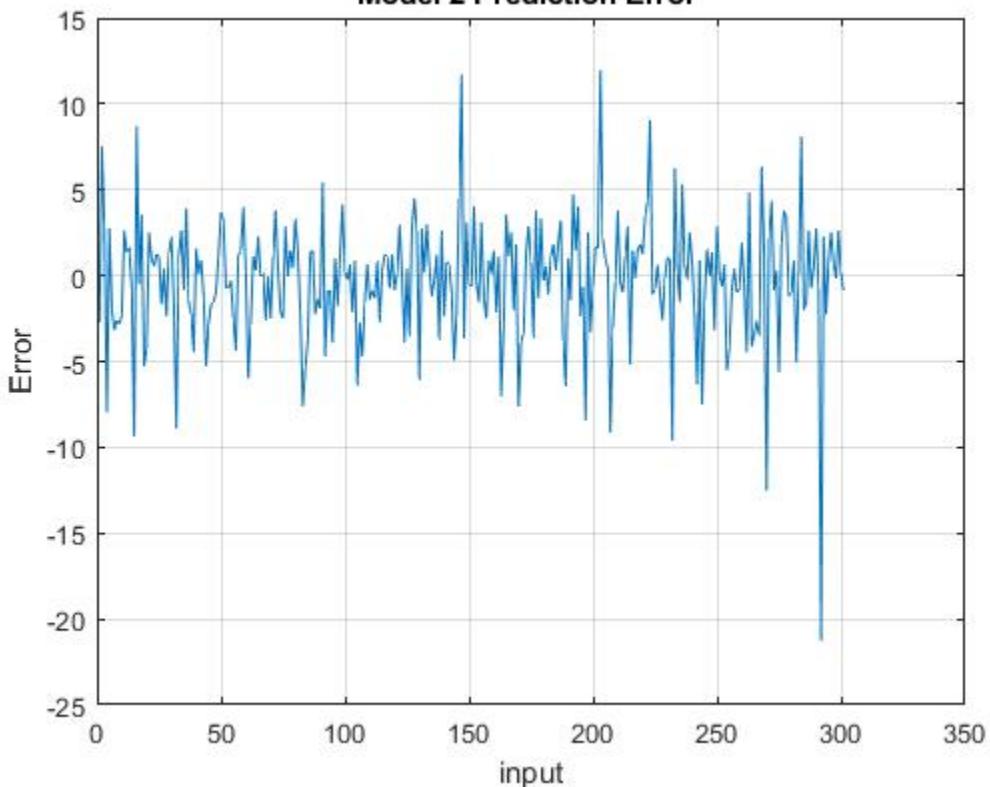
Παρατίθενται στη συνέχεια τα διαγράμματα μάθησης, όπου απεικονίζεται το σφάλμα του κάθε μοντέλου συναρτήσει του αριθμού των επαναλήψεων, καθώς και τα διαγράμματα όπου αποτυπώνονται τα σφάλματα πρόβλεψης του κάθε μοντέλου.

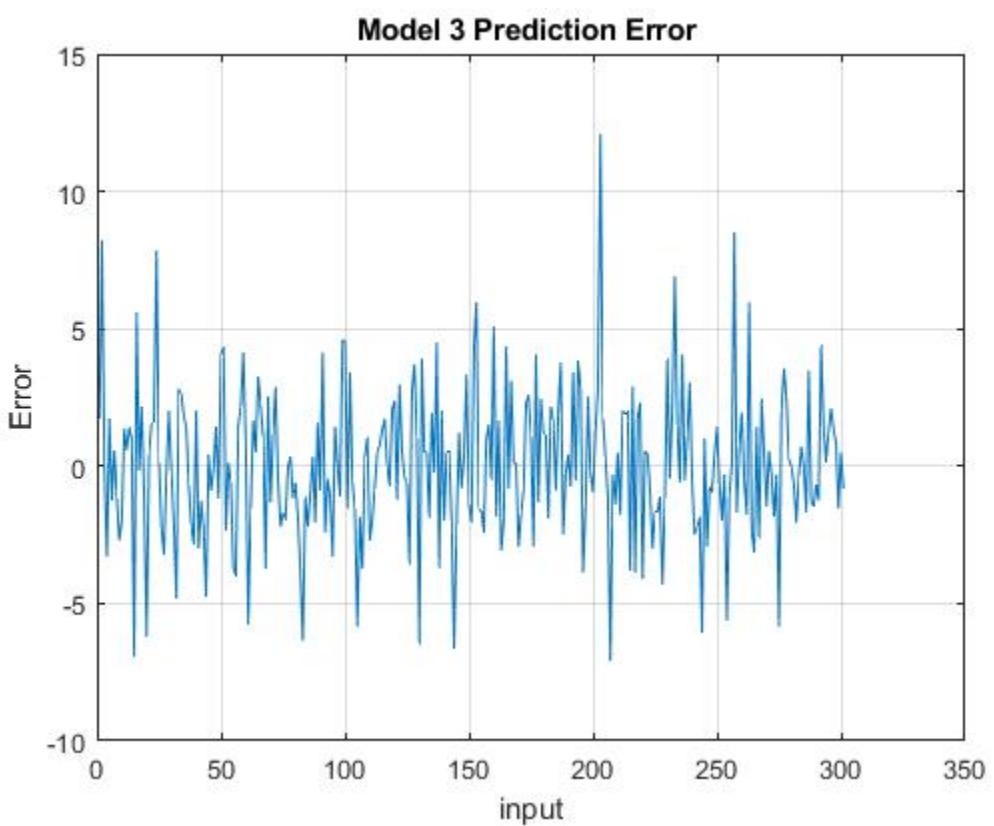
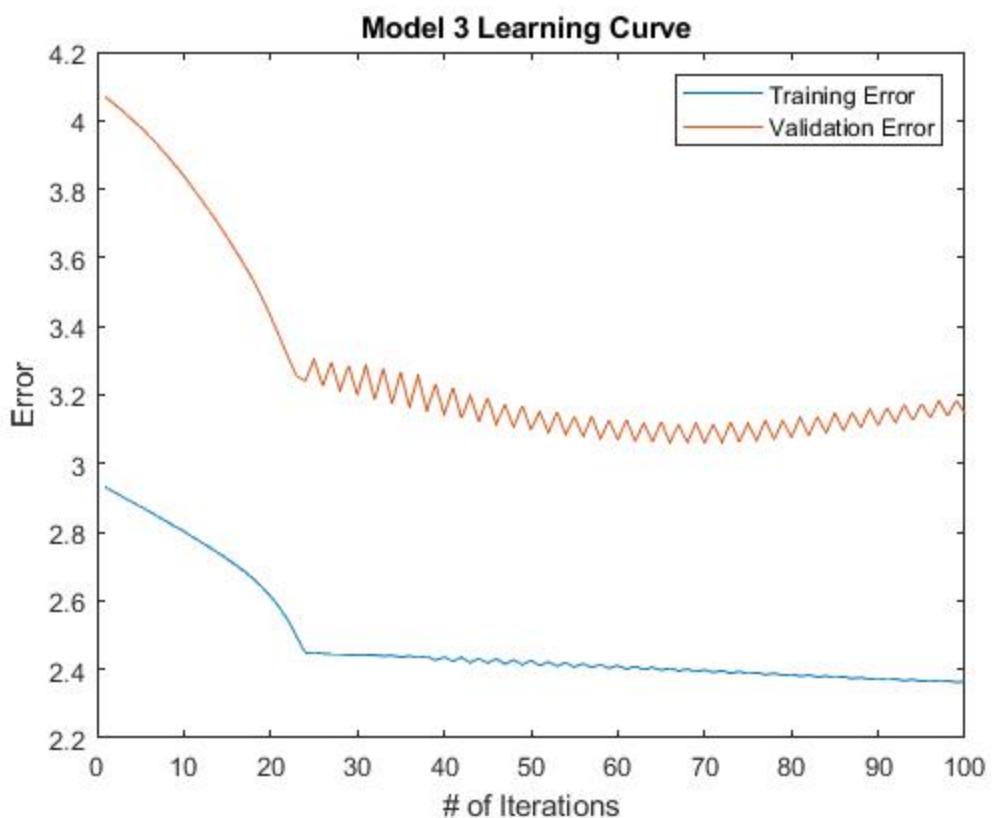


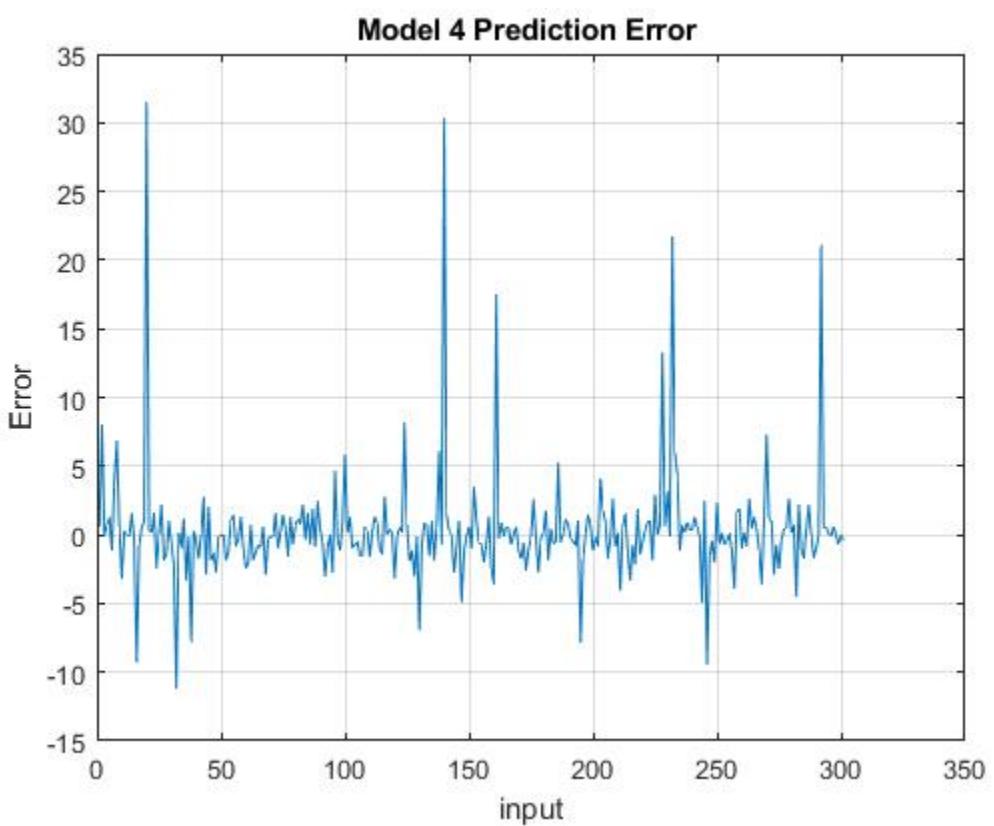
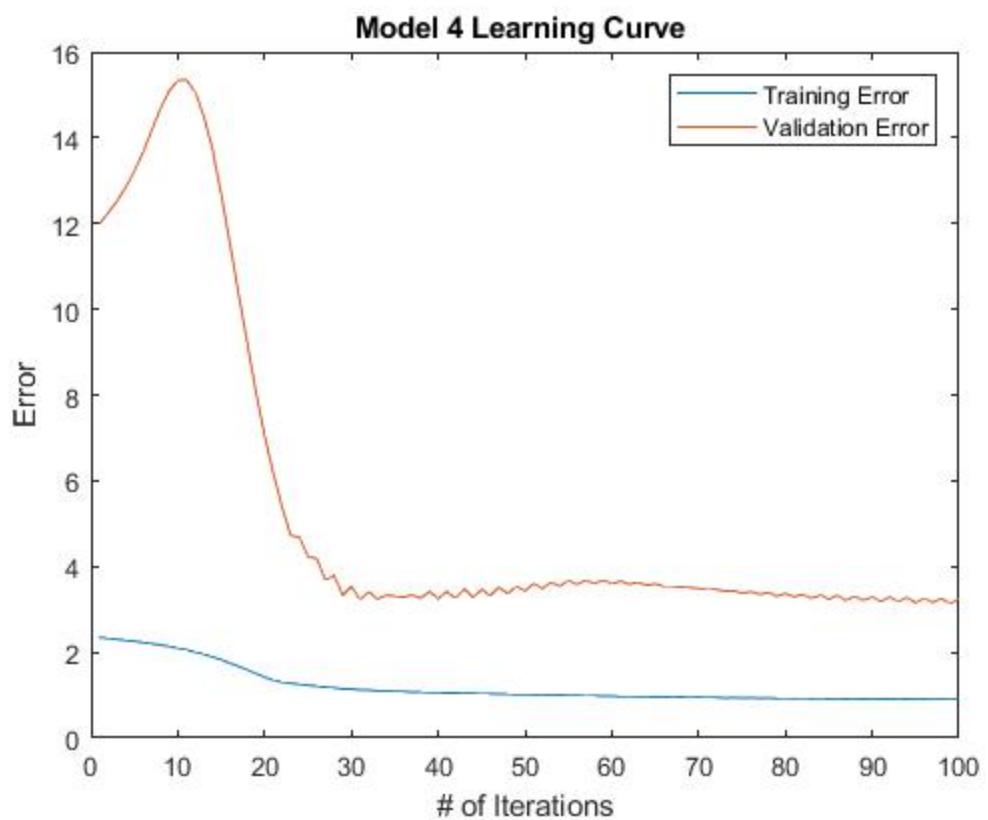
Model 2 Learning Curve



Model 2 Prediction Error







Παρατίθενται τέλος, οι τιμές των τεσσάρων δεικτών απόδοσης για κάθε μοντέλο.

	Model 1	Model 2	Model 3	Model 4
R²	0.6621	0.7192	0.8338	0.6340
RMSE	3.8397	3.5005	2.6932	3.9961
NMSE	0.3379	0.2808	0.1662	0.3660
NDEI	0.5813	0.5299	0.4077	0.6050

Βλέποντας τη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), συμπεραίνουμε ότι το μοντέλο 3 είναι το βέλτιστο. Συγκρίνοντας το 3ο με το 4ο μοντέλο, μπορούμε επίσης να συμπεράνουμε πως το μεγαλύτερο πλήθος ασαφών συνόλων ανά είσοδο, οδήγησε σε μεγαλύτερο RMSE, πιθανώς λόγω υπερεκπαίδευσης. Συνεπώς τα μοντέλα με περισσότερες συναρτήσεις συμμετοχής, αποκτούν περισσότερες πιθανότητες να οδηγηθούν σε υπερεκπαίδευση. Συγκρίνοντας εν συνεχείᾳ το μοντέλο 1 και το μοντέλο 3, τα οποία έχουν ίδιο αριθμό συναρτήσεων συμμετοχής, αλλά διαφορετική μορφή εξόδου, συμπεραίνουμε ότι η πολυωνυμική έξοδος τείνει να αποδίδει καλύτερα από τη singleton. Τα παραπάνω συμπεράσματα επιβεβαιώνονται από παραπάνω από μια εκτελέσεις, καθώς τα αποτελέσματα κάθε εκτέλεσης παρατηρήθηκε να διαφοροποιούνται αρκετά μεταξύ τους, και χρειάστηκαν αρκετές εκτελέσεις ώστε να εξάγουμε τα εν λόγω συμπεράσματα με μεγαλύτερη ασφάλεια.

Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Το δεύτερο μέρος της εργασίας υλοποιείται σε κώδικα MATLAB και αποθηκεύεται στο αρχείο main2.m. Σε πρώτη φάση, αφού φορτώνεται και κανονικοποιείται το σετ των δεδομένων που θα χρησιμοποιήσουμε, πρέπει να βρούμε βελτιστοποιημένες τιμές για τις δύο ελεύθερες παραμέτρους που περιλαμβάνει αναγκαστικά το πρόβλημα: τον αριθμό των χαρακτηριστικών προς επιλογή (features number) που θα χρησιμοποιηθούν στην εκπαίδευση των μοντέλων, και την ακτίνα επιρροής των clusters (r_a), η οποία καθορίζει το πλήθος των κανόνων που θα προκύψουν. Ο αριθμός των χαρακτηριστικών που θα εκπαιδεύουμε μπορεί να είναι από 1 μέχρι 80, ενώ η ακτίνα επιρροής παίρνει τιμές από 0 έως 1. Στο σημείο αυτό, πρέπει να επιλέξουμε τόσο τους διαφόρους συνδυασμούς των αριθμών των χαρακτηριστικών και των τιμών της ακτίνας, όσο και πόσους θέλουμε - συμφέρει να εκπαιδεύουμε και να ελέγχουμε σε κάθε εκτέλεση. Οι αριθμοί των χαρακτηριστικών και οι τιμές της ακτίνας για κάθε έναν από αυτούς, αποθηκεύονται σε δύο πίνακες (features_number, r_a _values), και πραγματοποιήθηκαν πολλαπλές εκτελέσεις για διάφορα μεγέθη του κάθε πίνακα, καθώς και για διαφόρους αριθμούς χαρακτηριστικών και τιμές της ακτίνας τη φορά. Το πρώτο συμπέρασμα που προέκυψε, είναι ότι από άποψη χρόνου εκτέλεσης και πολυπλοκότητας, δε συμφέρει η εκπαίδευση για πάνω από τρεις ή τέσσερις διαφορετικούς αριθμούς χαρακτηριστικών σε κάθε εκτέλεση, και ομοίως για τις διαφορετικές τιμές της ακτίνας. Εν συνεχεία, προέκυψε το συμπέρασμα ότι δε συμφέρει ο αριθμός των χαρακτηριστικών να ξεπερνάει τα 20-25, καθώς ο χρόνος εκτέλεσης αυξάνεται δραστικά, ενώ η απόδοση παρουσιάζει μικρές διαφορές συγκριτικά με μικρότερο αριθμό χαρακτηριστικών (όχι όμως πολύ μικρό). Παρατίθενται ενδεικτικά παρακάτω ορισμένα αποτελέσματα των δοκιμών που γίνανε, και συγκεκριμένα του δείκτη RMSE, ο οποίος αποτέλεσε και το βασικό κριτήριο για την επιλογή των δύο παραμέτρων που θα χρησιμοποιηθούν στο βέλτιστο, τελικό μοντέλο.

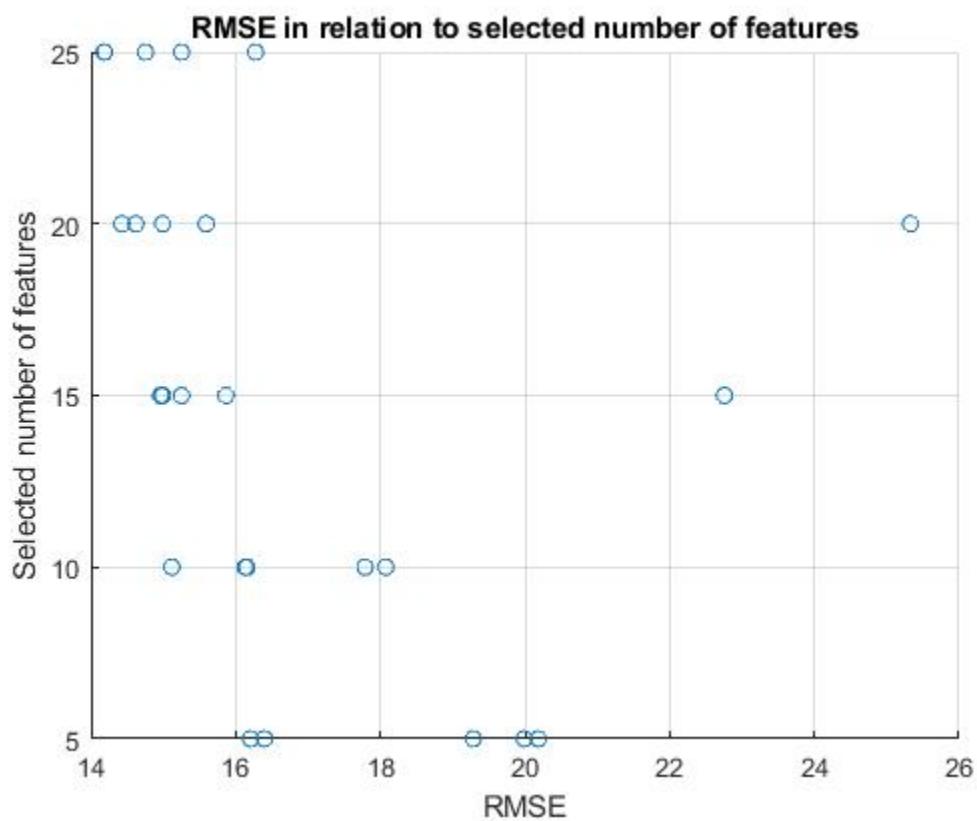
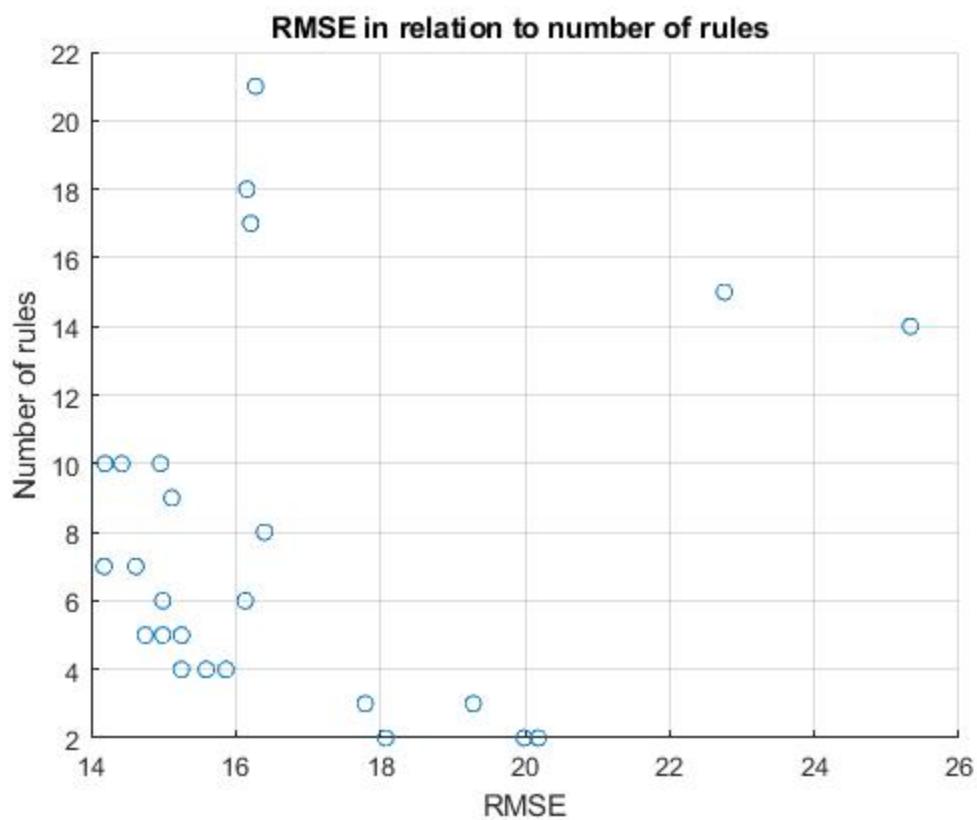
- Features = 5
 - $r_a = 0.2 \rightarrow RMSE = 16.2089$
 - $r_a = 0.25 \rightarrow RMSE = 16.6141$
 - $r_a = 0.3 \rightarrow RMSE = 16.0430$
 - $r_a = 0.4 \rightarrow RMSE = 16.3991$
 - $r_a = 0.5 \rightarrow RMSE = 17.0075$
 - $r_a = 0.6 \rightarrow RMSE = 19.0819$
 - $r_a = 0.75 \rightarrow RMSE = 20.0161$
 - $r_a = 0.8 \rightarrow RMSE = 20.1844$
 - $r_a = 1 \rightarrow RMSE = 19.9888$

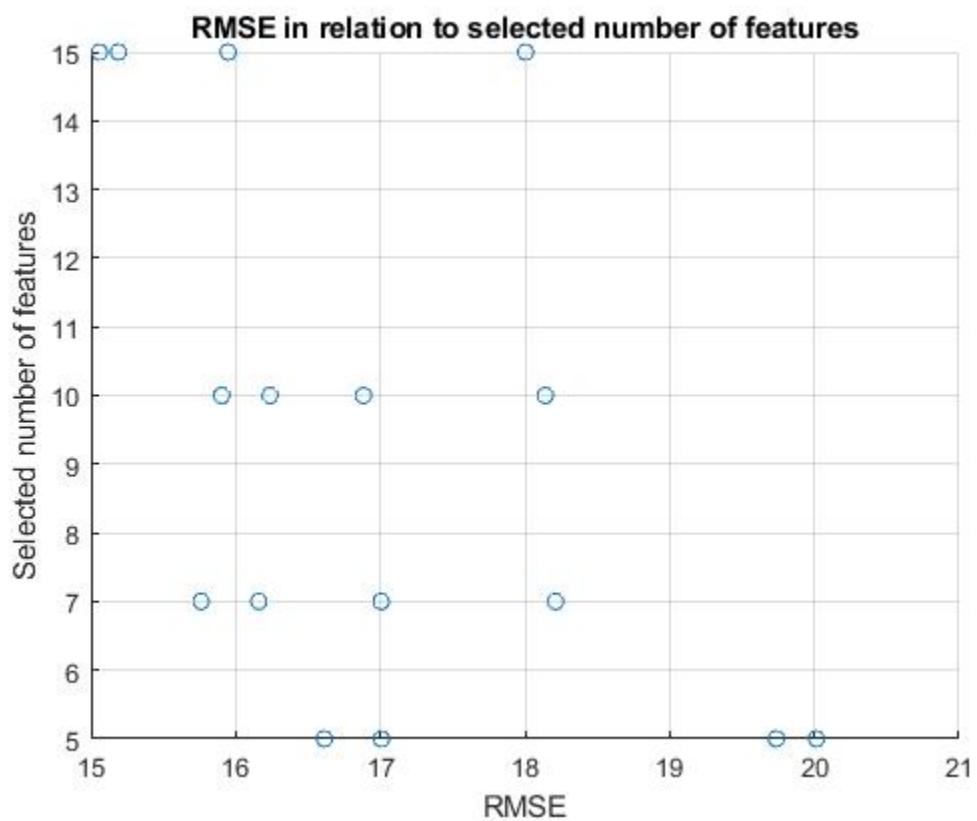
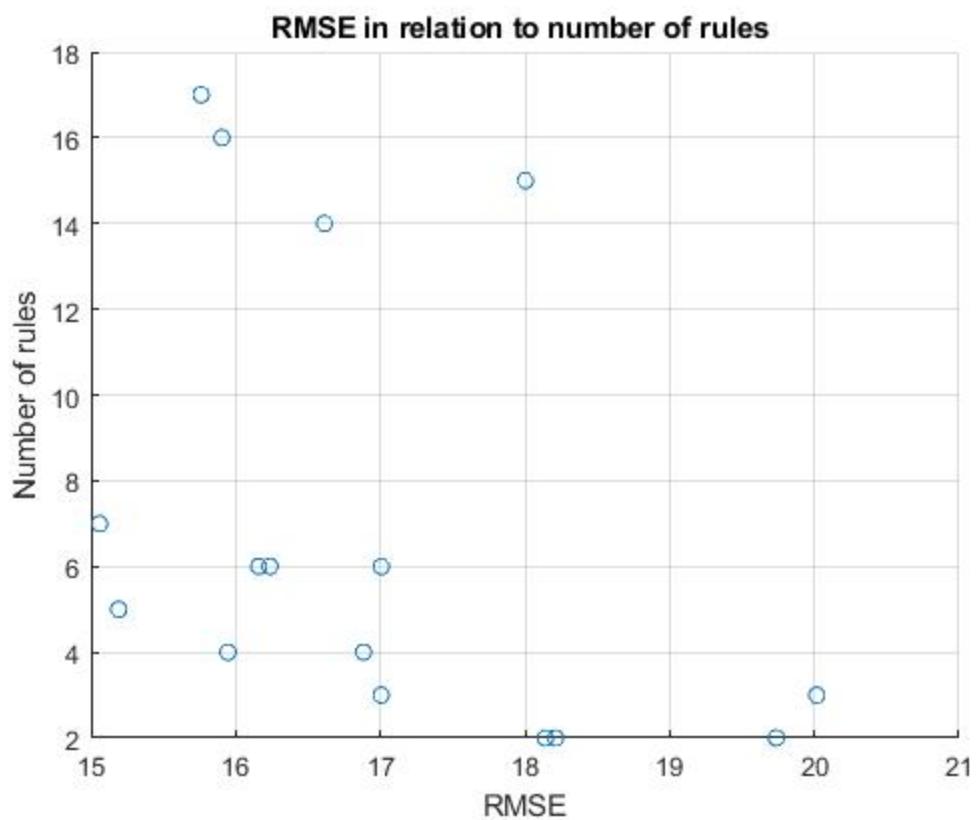
- Features = 7
 - $r_a = 0.2 \rightarrow$ not tested
 - $r_a = 0.25 \rightarrow$ RMSE = 15.7609
 - $r_a = 0.3 \rightarrow$ not tested
 - $r_a = 0.4 \rightarrow$ not tested
 - $r_a = 0.5 \rightarrow$ RMSE = 16.1591
 - $r_a = 0.6 \rightarrow$ not tested
 - $r_a = 0.75 \rightarrow$ RMSE = 17.0060
 - $r_a = 0.8 \rightarrow$ not tested
 - $r_a = 1 \rightarrow$ RMSE = 18.2121
- Features = 10
 - $r_a = 0.2 \rightarrow$ RMSE = 16.1566
 - $r_a = 0.25 \rightarrow$ RMSE = 15.9047
 - $r_a = 0.3 \rightarrow$ RMSE = 14.9957
 - $r_a = 0.4 \rightarrow$ RMSE = 15.1173
 - $r_a = 0.5 \rightarrow$ RMSE = 16.2377
 - $r_a = 0.6 \rightarrow$ RMSE = 16.1956
 - $r_a = 0.75 \rightarrow$ RMSE = 16.8841
 - $r_a = 0.8 \rightarrow$ RMSE = 17.7881
 - $r_a = 1 \rightarrow$ RMSE = 18.1069
- Features = 12
 - $r_a = 0.3 \rightarrow$ RMSE = 14.9077
 - $r_a = 0.4 \rightarrow$ RMSE = 14.7299
 - $r_a = 0.5 \rightarrow$ RMSE = 15.2624
- Features = 13
 - $r_a = 0.3 \rightarrow$ RMSE = 14.6049
 - $r_a = 0.4 \rightarrow$ RMSE = 14.5280
 - $r_a = 0.5 \rightarrow$ RMSE = 14.8996
- Features = 14
 - $r_a = 0.3 \rightarrow$ RMSE = 14.8241
 - $r_a = 0.4 \rightarrow$ RMSE = 14.6470

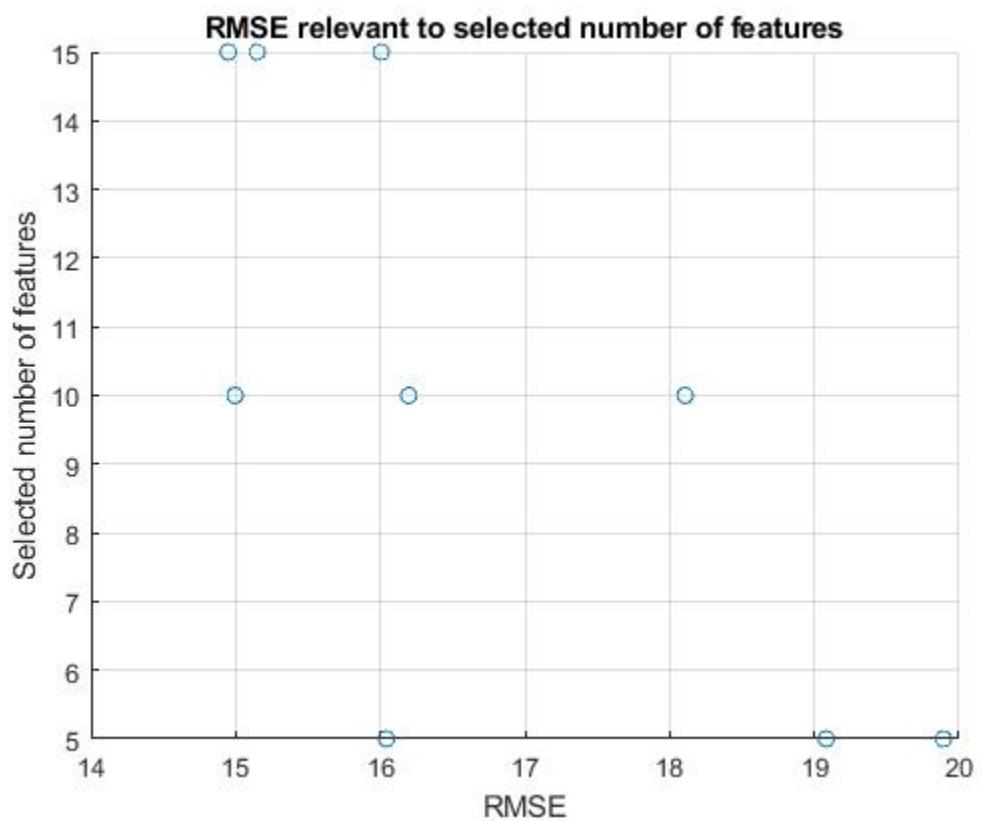
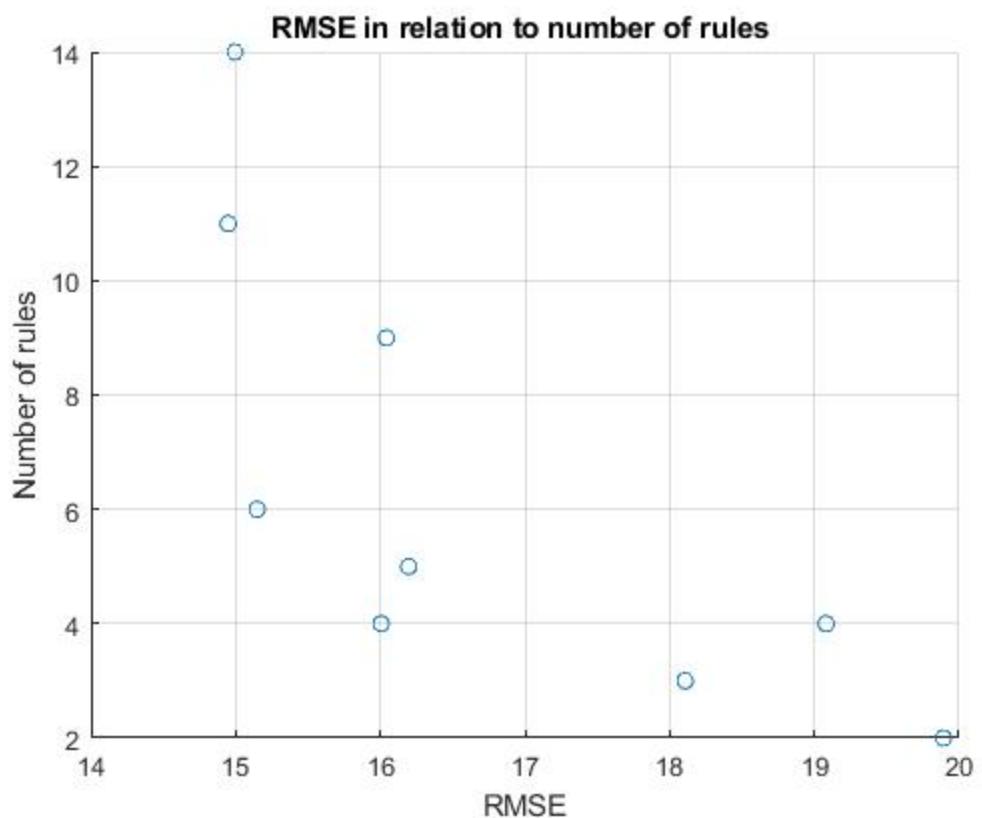
- $r_a = 0.5 \rightarrow \text{RMSE} = 14.7703$
- Features = 15
 - $r_a = 0.2 \rightarrow \text{RMSE} = 22.7568$
 - $r_a = 0.25 \rightarrow \text{RMSE} = 18.0038$
 - $r_a = 0.3 \rightarrow \text{RMSE} = 14.9484$
 - $r_a = 0.4 \rightarrow \text{RMSE} = 14.9607$
 - $r_a = 0.5 \rightarrow \text{RMSE} = 15.0587$
 - $r_a = 0.6 \rightarrow \text{RMSE} = 15.1479$
 - $r_a = 0.75 \rightarrow \text{RMSE} = 15.1912$
 - $r_a = 0.8 \rightarrow \text{RMSE} = 15.2536$
 - $r_a = 1 \rightarrow \text{RMSE} = 16.0053$
- Features = 20
 - $r_a = 0.2 \rightarrow \text{RMSE} = 25.3290$
 - $r_a = 0.25 \rightarrow \text{not tested}$
 - $r_a = 0.3 \rightarrow \text{not tested}$
 - $r_a = 0.4 \rightarrow \text{RMSE} = 14.4258$
 - $r_a = 0.5 \rightarrow \text{not tested}$
 - $r_a = 0.6 \rightarrow \text{RMSE} = 14.6204$
 - $r_a = 0.75 \rightarrow \text{not tested}$
 - $r_a = 0.8 \rightarrow \text{RMSE} = 14.9895$
 - $r_a = 1 \rightarrow \text{RMSE} = 15.5924$
- Features = 25
 - $r_a = 0.2 \rightarrow \text{RMSE} = 16.2733$
 - $r_a = 0.25 \rightarrow \text{not tested}$
 - $r_a = 0.3 \rightarrow \text{not tested}$
 - $r_a = 0.4 \rightarrow \text{RMSE} = 14.1914$
 - $r_a = 0.5 \rightarrow \text{not tested}$
 - $r_a = 0.6 \rightarrow \text{RMSE} = 14.1812$
 - $r_a = 0.75 \rightarrow \text{not tested}$
 - $r_a = 0.8 \rightarrow \text{RMSE} = 14.7514$
 - $r_a = 1 \rightarrow \text{RMSE} = 15.2505$

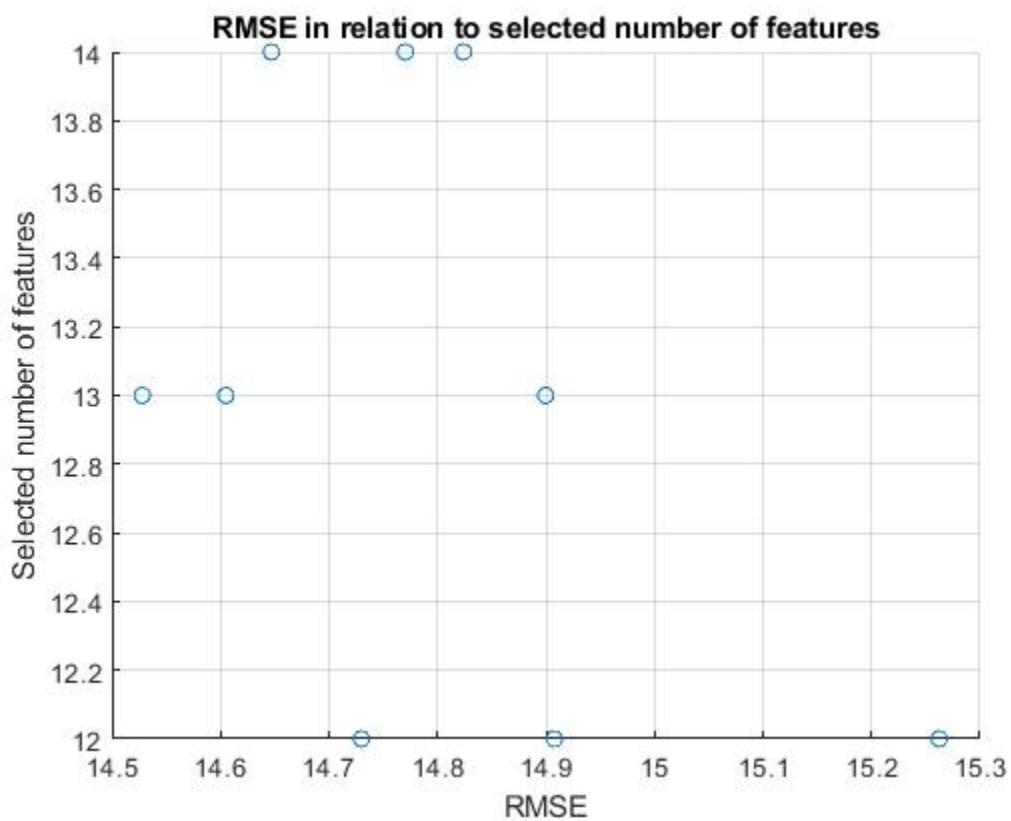
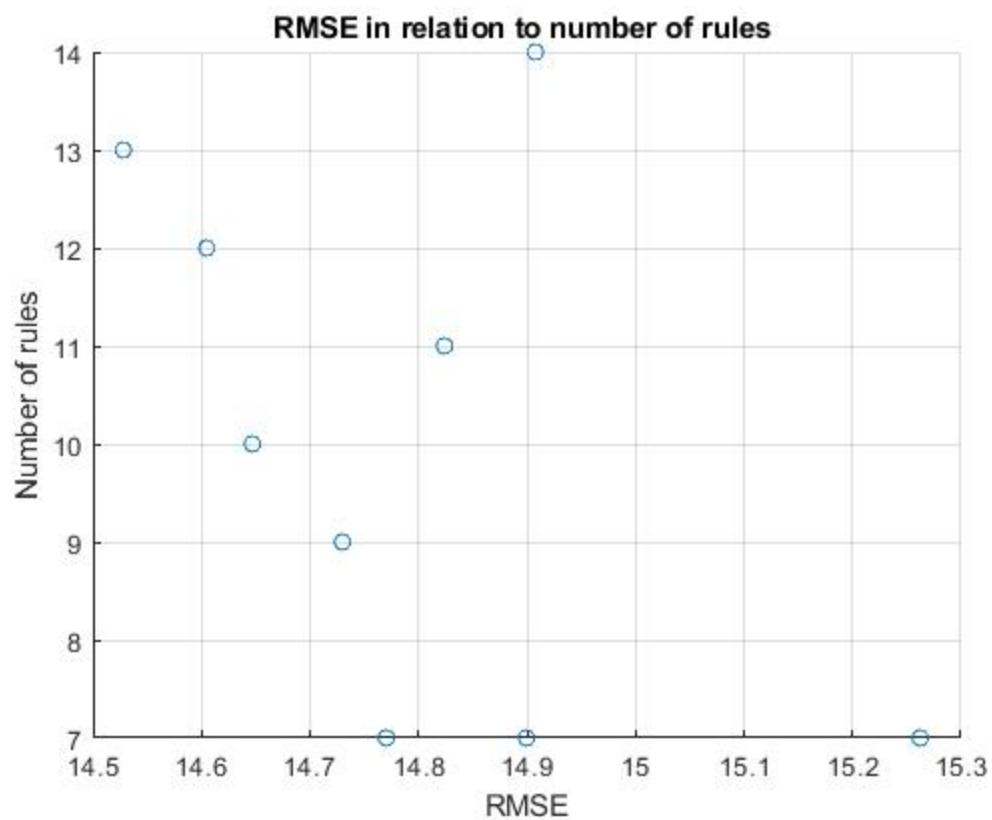
Παρατηρούμε πως παίρνουμε το μικρότερο σφάλμα όταν η ακτίνα έχει την τιμή 0.3 ή 0.4, και συγκεκριμένα για 13 χαρακτηριστικά ($RMSE = 14.52$). Βλέπουμε το σφάλμα να είναι ακόμα μικρότερο για 20 και 25 χαρακτηριστικά, ωστόσο όχι τόσο μικρότερο ώστε να αξίζει να χρησιμοποιήσουμε τόσο μεγάλο αριθμό χαρακτηριστικών στο τελικό μοντέλο, λόγω πολυπλοκότητας και χρόνου εκτέλεσης.

Παρατίθενται επίσης μερικά διαγράμματα πολλαπλών εκτελέσεων και δοκιμών, τα οποία απεικονίζουν το σφάλμα ($RMSE$) συναρτήσει του αριθμού των κανόνων και του αριθμού των επιλεχθέντων χαρακτηριστικών.









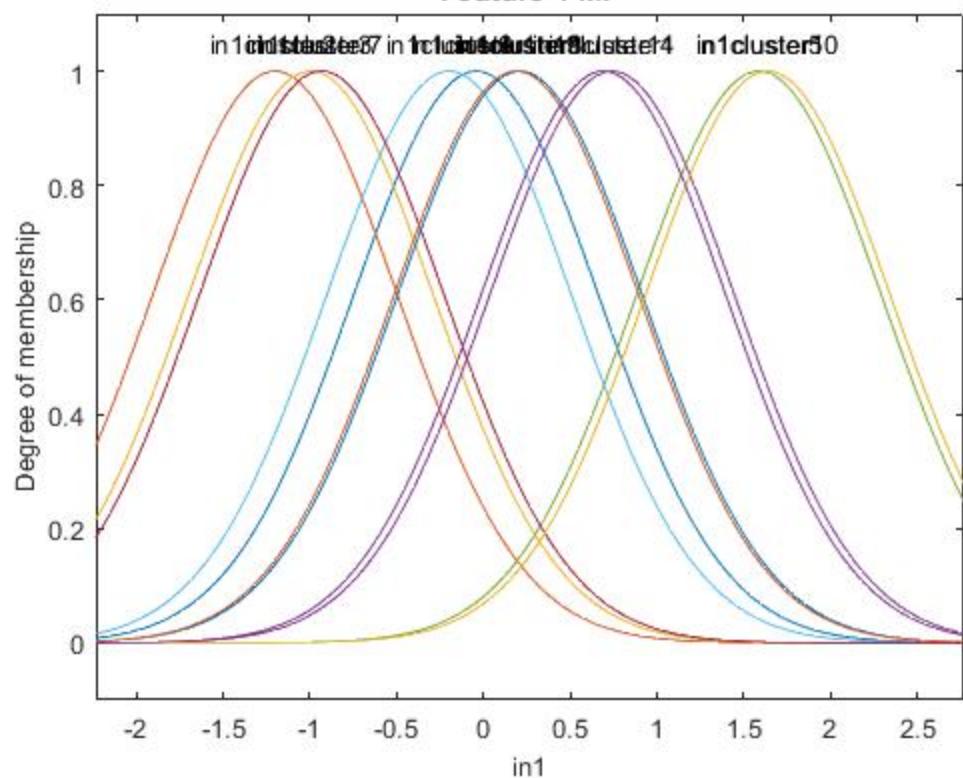
Για κάθε διαφορετικό αριθμό χαρακτηριστικών, και για κάθε διαφορετική τιμή της ακτίνας, διαχωρίζουμε τα δεδομένα σε ποσοστό 80% για εκπαίδευση και 20% για έλεγχο, και ορίζουμε έναν πίνακα για να αποθηκεύσουμε τους ζητούμενους δείκτες απόδοσης (τέσσερις στο σύνολο) που θα προκύψουν μετά το cross validation. Εν συνέχεια προχωρούμε στο cross validation, το οποίο είναι 5-fold, και μέσα στο οποίο διαχωρίζω το 80% των δεδομένων εκπαίδευσης από προηγουμένως, σε ποσοστό 75% για εκπαίδευση, και 25% για επικύρωση. Έτσι, το αρχικό σετ δεδομένων έχει διαχωριστεί σε ποσοστό 60% δεδομένα εκπαίδευσης (trnData), 20% δεδομένα επικύρωσης, και 20% δεδομένα ελέγχου (tstData), όπως ορίζει η εκφώνηση. Εν συνεχεία εκπαιδεύονται πέντε διαφορετικά μοντέλα με τις ίδιες παραμέτρους (5-fold cross validation), για 100 epochs το καθένα, και υπολογίζονται οι τέσσερις ζητούμενοι δείκτες απόδοσης του καθενός, οι οποίοι αποθηκεύονται στη συνέχεια στον πίνακα που ορίσαμε στην αρχή. Τέλος, μετά το πέρας του cross validation, υπολογίζουμε το μέσο όρο του κάθε δείκτη, ο οποίος έχει λάβει πέντε διαφορετικές τιμές κατά τη διάρκεια, και τον αποθηκεύουμε στον τελικό πίνακα όπου θα έχουμε όλους τους δείκτες για κάθε συνδυασμό αριθμού χαρακτηριστικών και τιμών της ακτίνας που επιλέξαμε στην αρχή.

Εκπαίδευση του τελικού, βέλτιστου μοντέλου

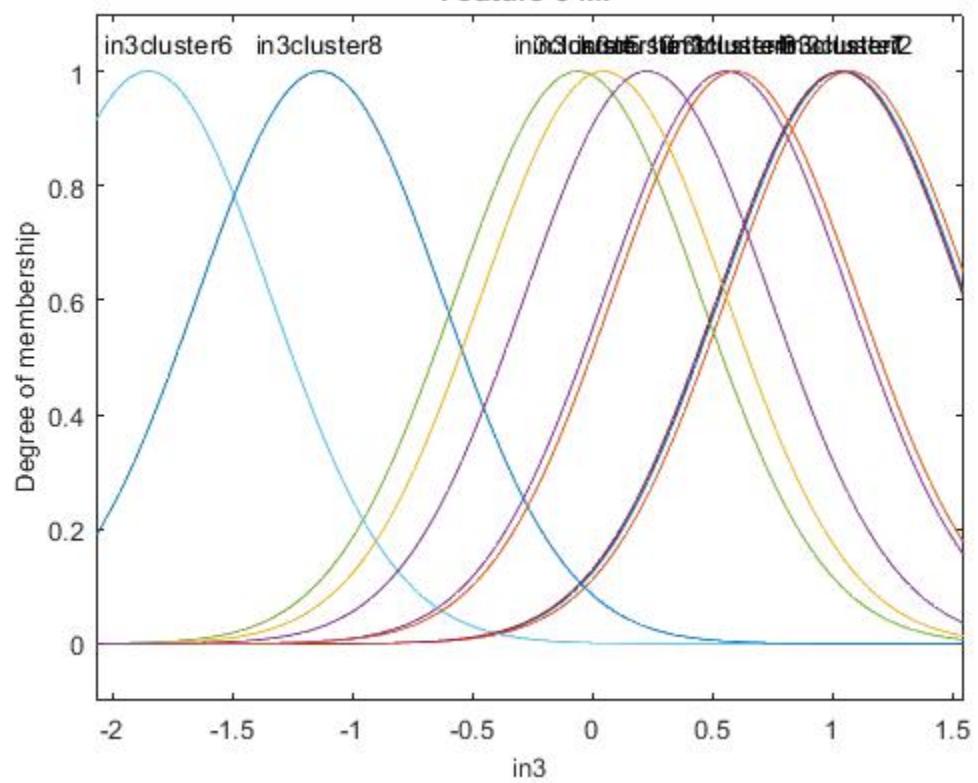
Το τελευταίο μέρος της εργασίας υλοποιείται σε κώδικα MATLAB και αποθηκεύεται στο αρχείο main3.m, όπου και εκπαιδεύομε το τελικό, βέλτιστο μοντέλο χρησιμοποιώντας τον αριθμό των χαρακτηριστικών και την τιμή της ακτίνας που επιλέχθηκαν βάση των συμπερασμάτων του δευτέρου μέρους της εργασίας. Ο αριθμός των χαρακτηριστικών που επιλέχθηκε είναι 13, ενώ η τιμή της ακτίνας είναι 0.4. Το μοντέλο εκπαιδεύεται ακριβώς με τον ίδιο τρόπο όπως τα προηγούμενα, στη συνέχεια υπολογίζουμε και αποθηκεύουμε το μέσο όρο των τεσσάρων δεικτών απόδοσης που μας ενδιαφέρουν και προκύπτουν από τη διαδικασία του cross validation, και δημιουργούμε ενδεικτικά μερικά διαγράμματα συναρτήσεων συμμετοχής στην αρχική και στην τελική τους μορφή, τα διαγράμματα όπου αποτυπώνονται οι προβλέψεις του τελικού μοντέλου και οι πραγματικές τιμές, το διάγραμμα εκμάθησης, όπου απεικονίζεται το σφάλμα συναρτήσει του αριθμού των επαναλήψεων, και τα οποία παρατίθενται παρακάτω, μαζί με τον πίνακα των τιμών των τεσσάρων δεικτών απόδοσης.

Συναρτήσεις συμμετοχής για διαφορετικά χαρακτηριστικά πριν την εκπαίδευση

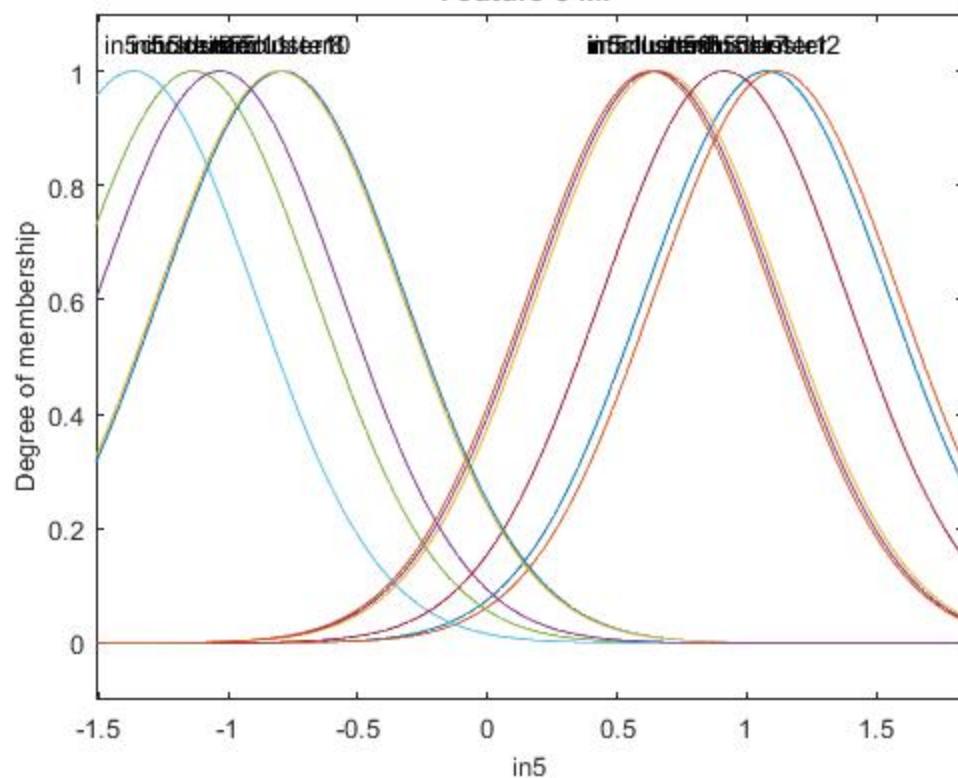
Feature 1 MF



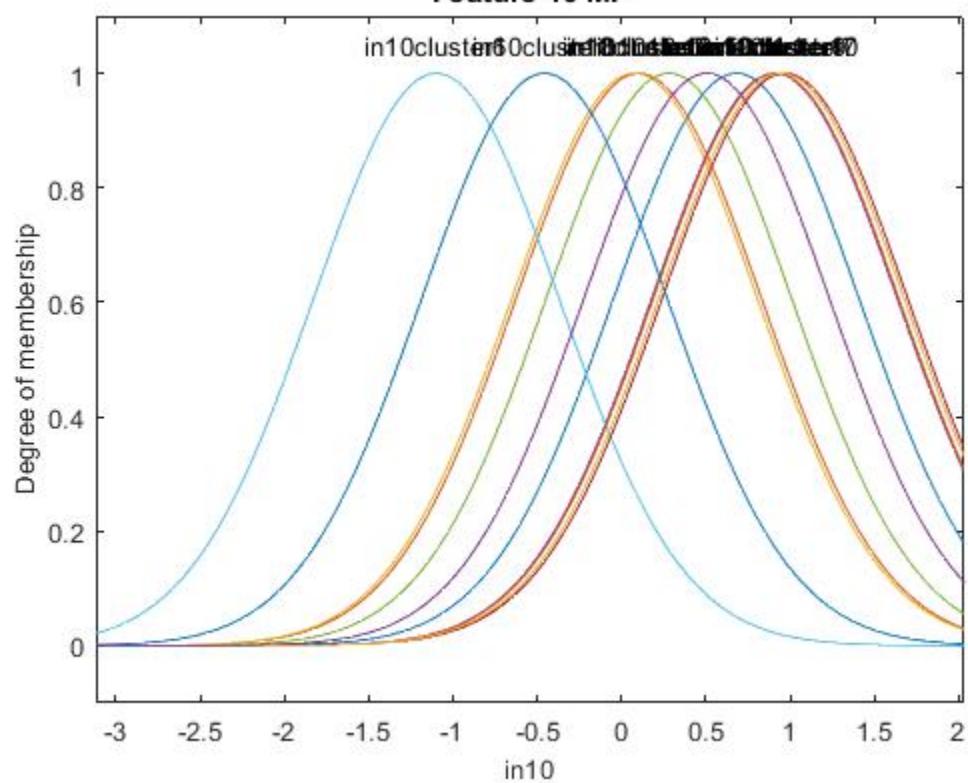
Feature 3 MF



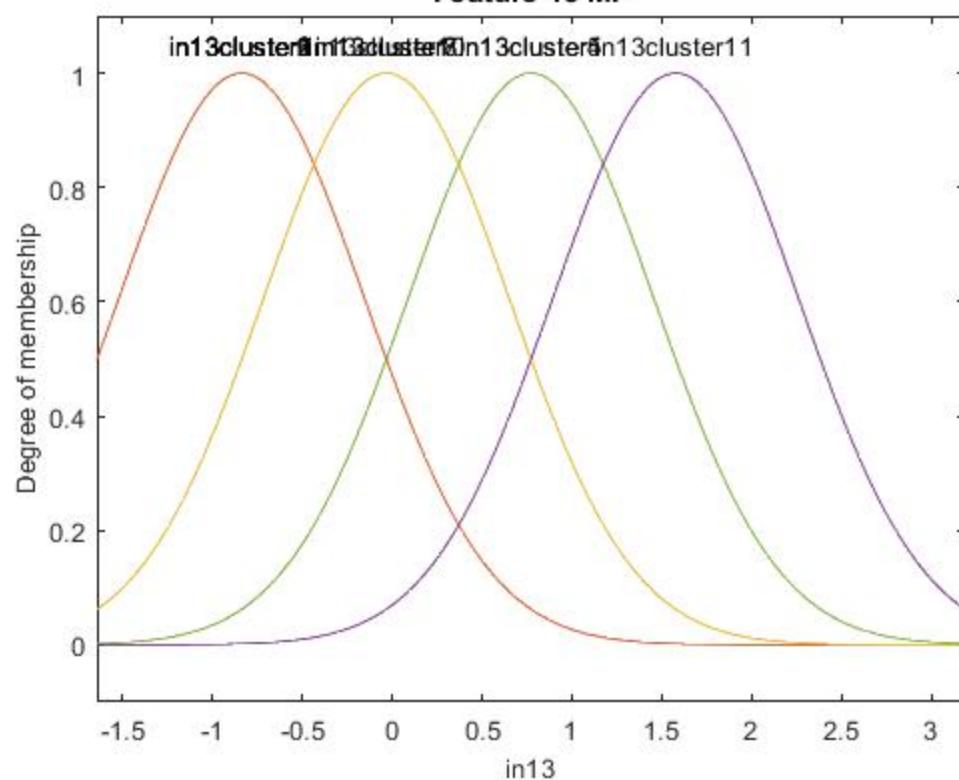
Feature 5 MF



Feature 10 MF

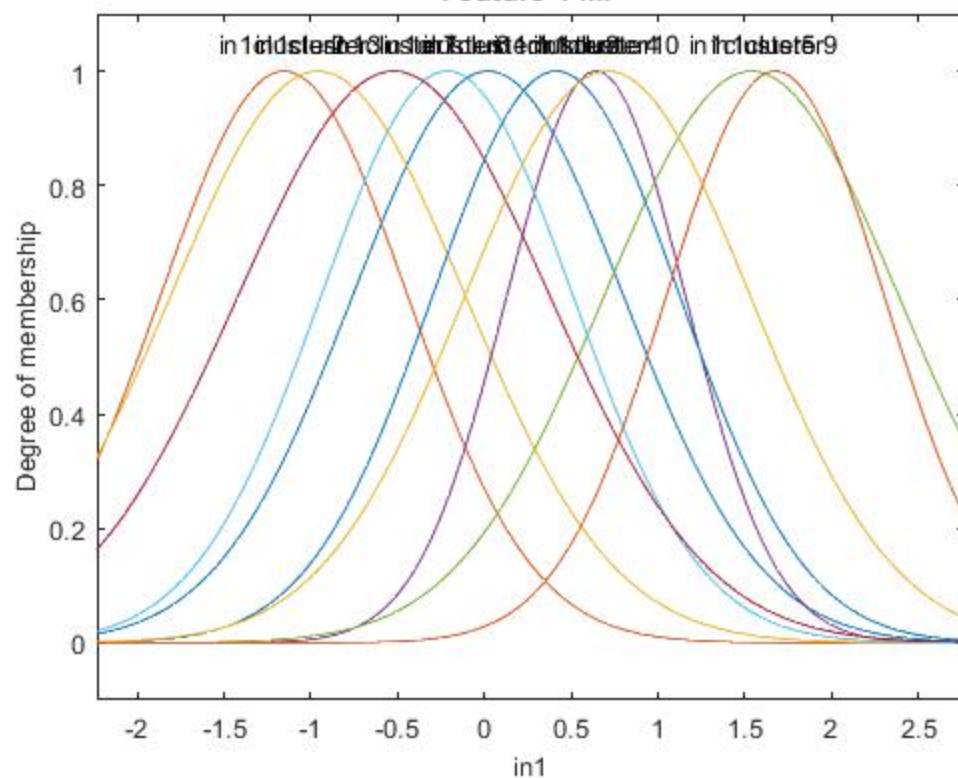


Feature 13 MF

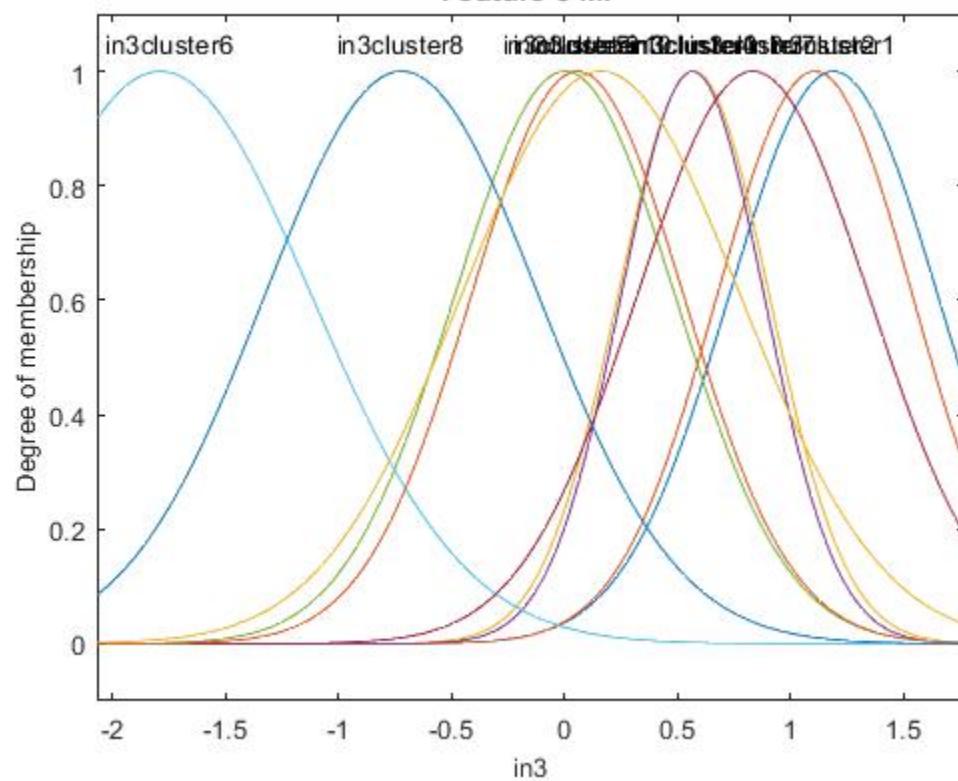


Συναρτήσεις συμμετοχής για τα ίδια χαρακτηριστικά μετά την εκπαίδευση

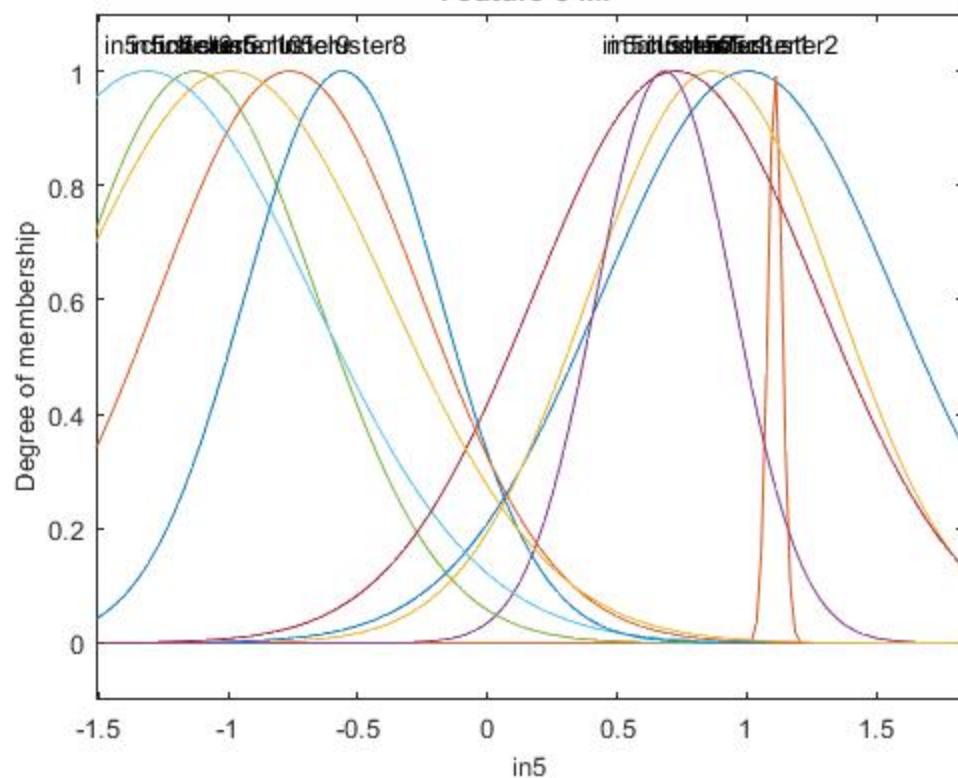
Feature 1 MF



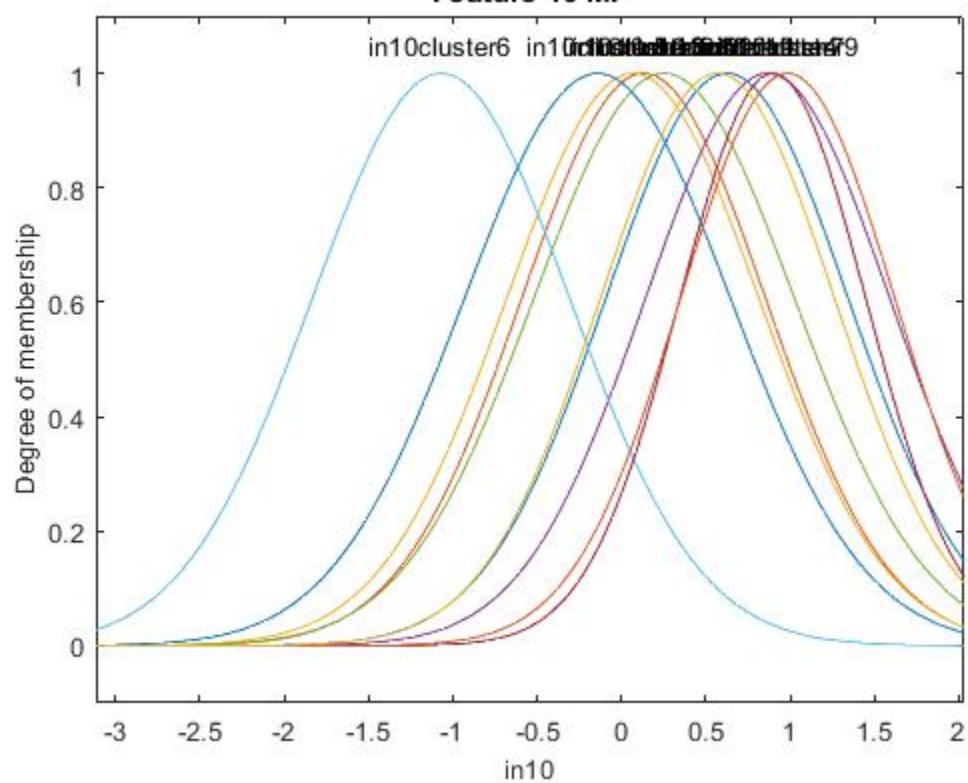
Feature 3 MF



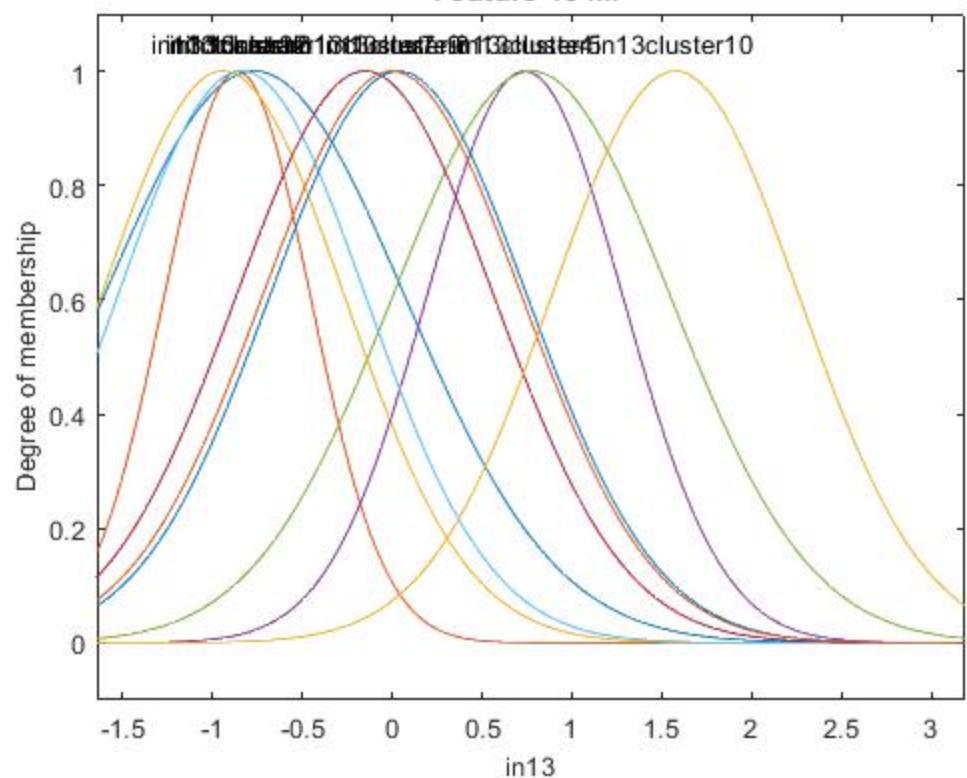
Feature 5 MF



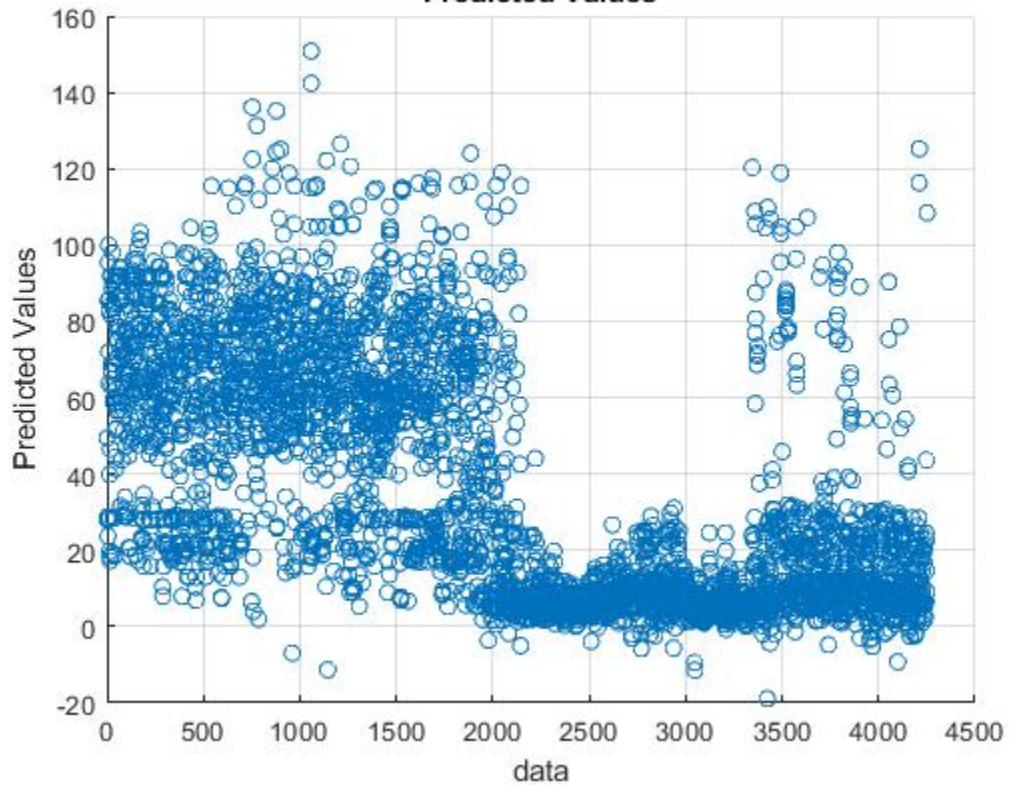
Feature 10 MF

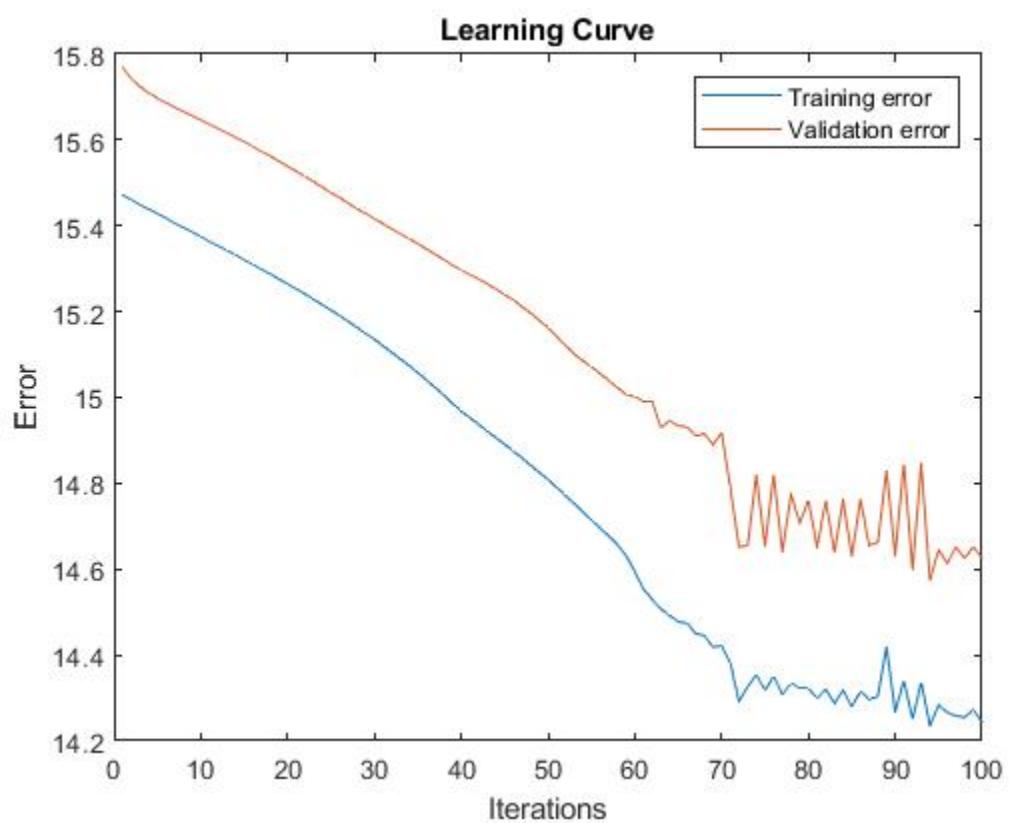
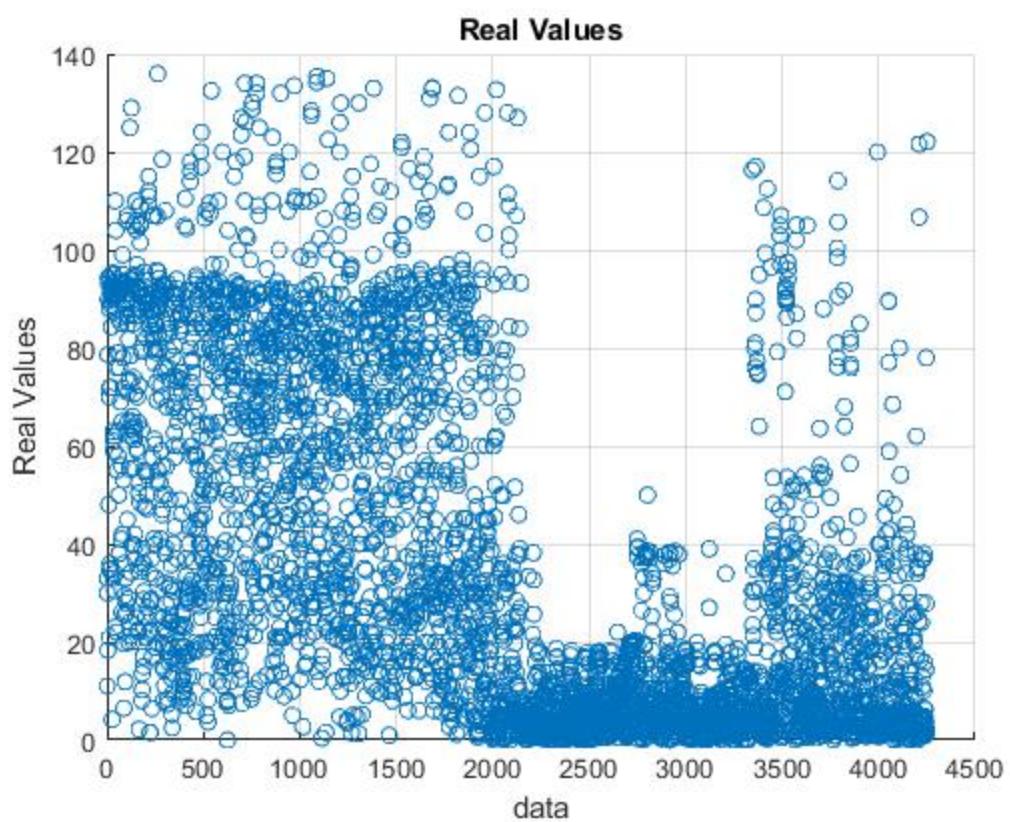


Feature 13 MF



Predicted Values





R²	0.8004
RMSE	15.2760
NMSE	0.1996
NDEI	0.4460

Ο αριθμός των κανόνων του ασαφούς συστήματος συμπερασμού είναι στην προκειμένη περίπτωση 10. Εάν για το ίδιο πλήθος χαρακτηριστικών είχαμε επιλέξει grid partitioning με δύο ή τρία ασαφή σύνολα ανά είσοδο, ο εν λόγω αριθμός των κανόνων θα εκτοξευόταν σε 2^{13} και 3^{13} αντίστοιχα. Το γεγονός αυτό καθιστά όχι μονο απαγορευτική την εφαρμογή και την εκτέλεσή της πρακτικά, αλλά θα είχε και πολύ μεγάλες πιθανότητες να εμφανίσει υπερεκπαίδευση.