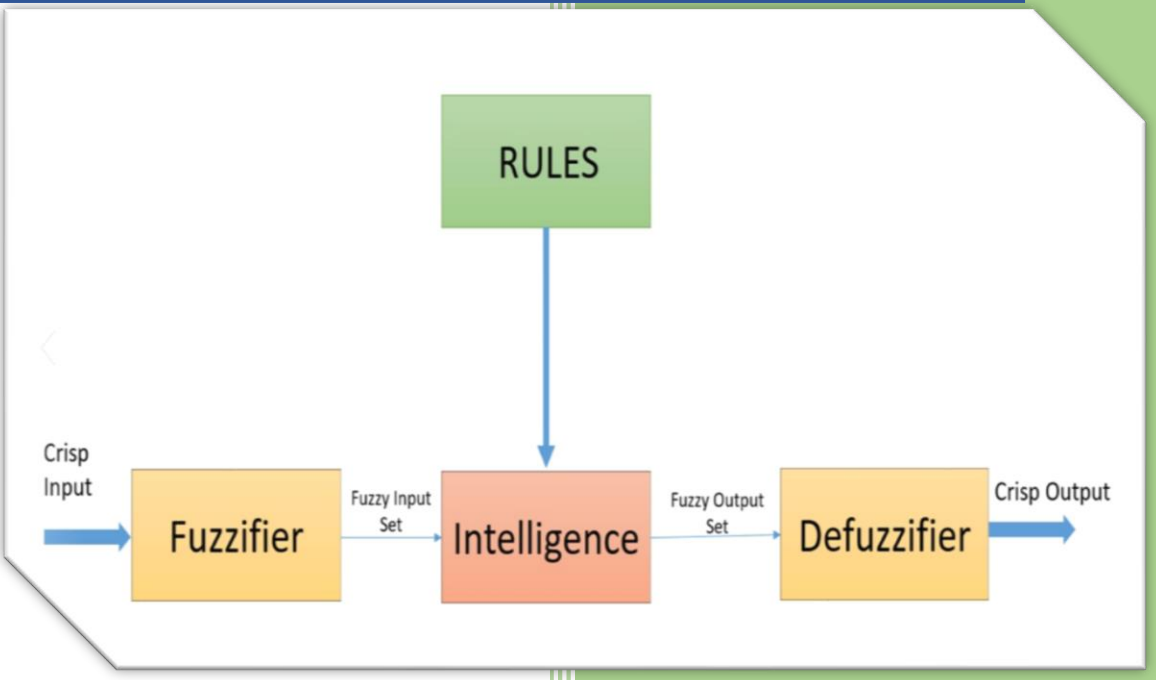
2019

Ασαφή Συστήματα – Εργασία IV



Τογκουσίδης Αναστάσιος

ΑΕΜ: 8920

25/9/2019

Πρώτο μέρος

Η τέταρτη εργασία που μου ανατέθηκε είναι η **Group 4 Set 2**. Η εργασία διερευνά την ικανότητα των μοντέλων TSK στην επίλυση των προβλημάτων ταξινόμησης. Το πρώτο μέρος της εργασίας αποτελεί μια απλή διερεύνηση της διαδικασίας εκπαίδευσης και αξιολόγησης των μοντέλων. Για το σκοπό αυτό επιλέγεται από το UCI Repository το avila dataset που αποτελείται από 20876 δείγματα (instances) με δέκα χαρακτηριστικά (attributes) το καθένα.

Η μεταβλητή που θεωρείται έξοδος στο σύστημα είναι ο αριθμός της κλάσης που ανήκει το κάθε δεδομένο. 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. Επίσης, προκειμένου να επιτευχθεί ικανοποιητική απόδοση, τα δεδομένα διαχωρίζονται με τέτοιο τρόπο, ώστε η συχνότητα εμφάνισης των δειγμάτων που ανήκουν σε μια συγκεκριμένη κλάση, σε κάθε ένα από τα σύνολα διαμέρισης, να είναι το κατά δύναμιν όμοια με την αντίστοιχη συγχότητα εμφάνισής τους στο αρχικό σύνολο δεδομένων.

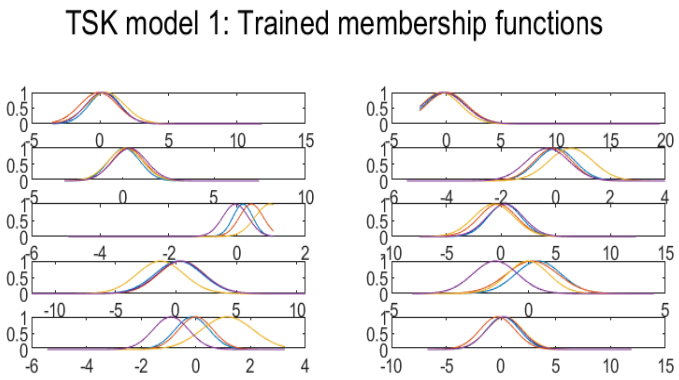
Δυστυχώς, το σύνολο που μας δόθηκε να χρησιμοποιήσουμε δεν μας δίνει ικανοποιητικά αποτελέσματα (σύμφωνα με παρατήρηση του κυρίου Χαδουλού σε απάντηση email σε συμφοιτητή μου). Ως εκ τούτου, τα αποτελέσματα που θα παρουσιαστούν είναι ιδιαίτερα απογοητευτικά.

Ζητείται να δημιουργηθούν μοντέλα στα οποία το σύνολο των κανόνων να λαμβάνει τιμές στο σύνολο $NR = \{4, 8, 12, 16, 20\}$. Οι συγκεκριμένοι κανόνες προέκυπταν από πολύ μικρές τιμές στην παράμετρο randii, το οποίο είχε σαν αποτέλεσμα να αδυνατεί το Matlab να κάνει υπολογισμούς. Χρησιμοποίησα, λοιπόν, το σύνολο $NR = \{2,3,4,5,6\}$. Οι αντίστοιχες τιμές randii που βρέθηκαν από τη συνάρτηση `find_centers.m` στον κώδικα του Matlab είναι $[0.24, 0.18, 0.16, 0.145, 0.13]$. Και πάλι, ο κώδικας έδωσε αποτελέσματα μόνο για τις πρώτες τρεις τιμές του randii.

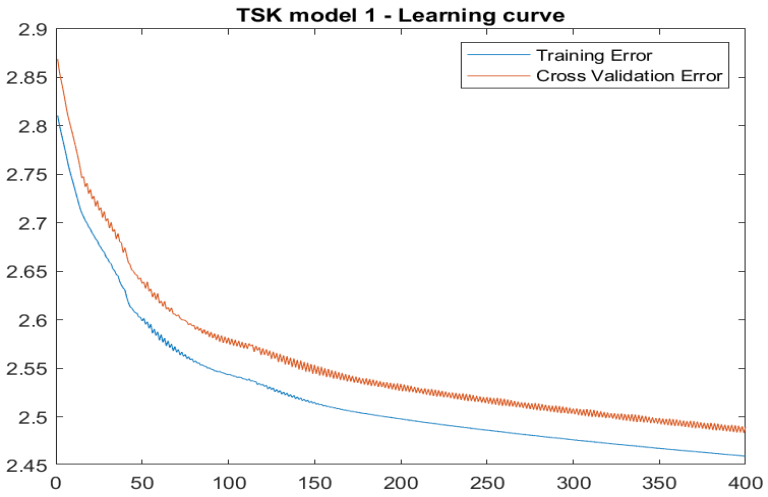
Για τη δημιουργία των μοντέλων χρησιμοποιείται η συνάρτηση `genfis2()` του Matlab. Ο λόγος που χρησιμοποιείται η συγκεκριμένη συνάρτηση είναι επειδή μας επιτρέπει να περάσουμε εύκολα τις παραμέτρους που θέλουμε στη δημιουργία του κάθε μοντέλου και πραγματοποιεί αυτόματα την τεχνική Subtractive Clustering. 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). Παρακάτω παρουσιάζονται τα αποτελέσματα για $NR = 2, 3, 4$.

NR = 2

Στα σχήματα 1 και 2 παρουσιάζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν από τη διαδικασία εκπαίδευσης και τα διαγράμματα μάθησης.



Σχήμα 1: Τελικές μορφές ασαφών συνόλων



Σχήμα 2: Καμπύλη εκμάθησης

Στην επόμενη σελίδα παρουσιάζονται ο πίανακας σφαλμάτων και εξάγονται απ’ αυτόν οι τιμές των δεικτών απόδοσης.

79	0	0	9	1	7	0	0	0	4	0	0
306	0	0	20	8	51	3	2	0	6	1	0
789	0	16	61	64	271	27	25	14	2	12	6
404	0	11	22	110	293	62	47	15	4	10	7
102	2	7	19	111	113	57	39	11	0	5	3
20	0	5	6	92	36	22	51	2	0	11	1
6	0	3	2	37	6	6	33	3	0	8	7
1	0	0	0	9	0	1	10	4	0	18	4
2	0	0	1	6	0	1	1	17	1	32	13
2	0	0	1	0	0	0	0	267	0	112	66
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 1: Πίνακας σφαλμάτων

$$OA = 0.0711$$

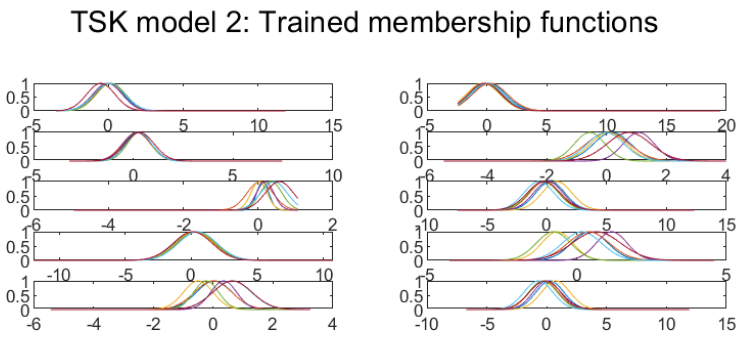
$$PA = [0.046, 0, 0.38, 0.156, 0.2534, 0.046, 0.033, 0.048, 0.051, 0, 0, 0]$$

$$U = [0.79, 0, 0.012, 0.0223, 0.236, 0.146, 0.054, 0.212, 0.229, 0, NaN, NaN]$$

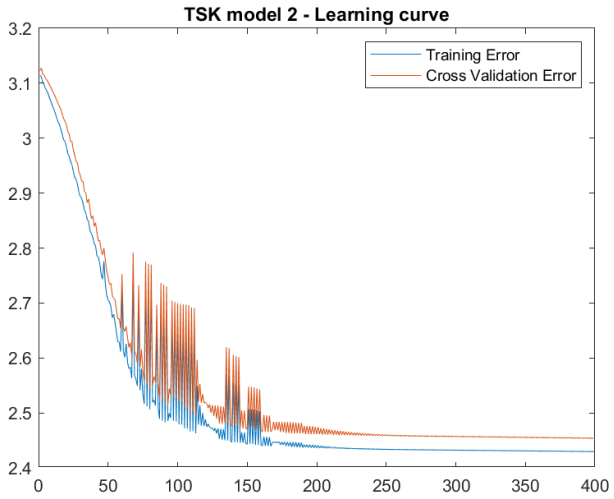
$$\hat{\kappa} = 0.0015$$

NR = 3

Στα σχήματα 3 και 4 παρουσιάζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν από τη διαδικασία εκπαίδευσης και τα διαγράμματα μάθησης.



Σχήμα 3: Τελικές μορφές ασαφών συνόλων



Σχήμα 4: Καμπύλη εκμάθησης

Στην επόμενη σελίδα παρουσιάζονται ο πίανακας σφαλμάτων και εξάγονται απ' αυτόν οι τιμές των δεικτών απόδοσης.

68	0	0	2	2	2	1	0	0	2	0	0
349	0	1	43	6	79	2	5	0	4	0	0
797	2	13	57	79	304	37	23	5	6	4	9
301	0	7	11	73	203	51	28	1	0	10	3
147	0	13	16	99	149	49	38	7	0	4	2
43	0	6	8	95	47	33	63	38	0	15	2
5	0	2	2	48	1	4	39	11	3	13	7
2	0	0	2	26	0	0	12	50	0	20	8
0	0	0	0	9	0	2	0	79	2	23	6
0	0	0	0	1	0	0	0	93	1	63	44
0	0	0	0	0	0	0	0	49	0	57	26
0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 2: Πίνακας σφαλμάτων

$$OA = 0.0936$$

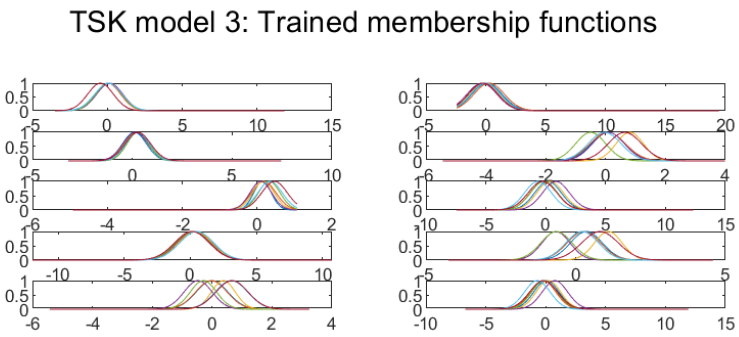
$$PA = [0.039, 0, 0.309, 0.078, 0.226, 0.059, 0.022, 0.057, 0.237, 0.055, 0.272, 0]$$

$$U = [0.883, 0, 0.009, 0.016, 0.188, 0.134, 0.029, 0.1, 0.652, 0.005, 0.431, NaN]$$

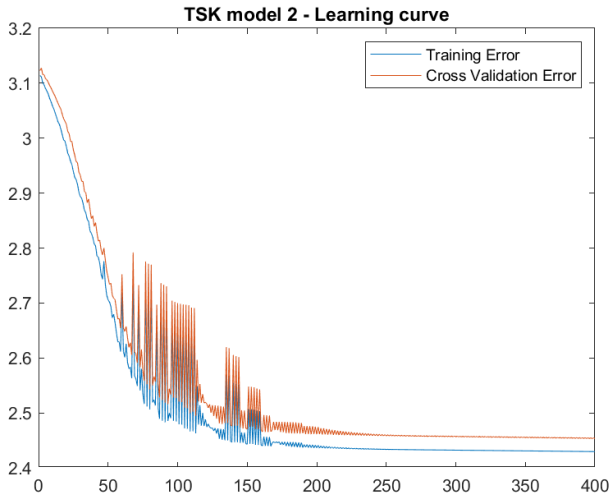
$$\hat{\kappa} = 0.0254$$

NR = 4

Στα σχήματα 5 και 6 παρουσιάζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν από τη διαδικασία εκπαίδευσης και τα διαγράμματα μάθησης.



Σχήμα 5: Τελικές μορφές ασαφών συνόλων



Σχήμα 6: Καμπύλη εκμάθησης

Στην επόμενη σελίδα παρουσιάζονται ο πίανακας σφαλμάτων και εξάγονται απ' αυτόν οι τιμές των δεικτών απόδοσης.

56	0	0	0	0	5	0	0	0	3	0	0
289	0	0	22	5	50	3	5	0	4	0	0
806	2	12	69	70	337	26	23	8	5	8	15
329	0	7	16	64	187	56	25	10	0	4	9
145	0	11	17	91	165	58	51	16	0	6	4
67	0	9	12	97	34	23	46	52	1	17	3
4	0	3	2	57	1	10	41	22	1	9	4
5	0	0	3	28	0	3	17	79	0	21	7
1	0	0	0	20	0	0	0	47	0	26	19
1	0	0	0	6	0	0	0	99	4	118	46
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Πίνακας 3: Πίνακας σφαλμάτων

$$OA = 0.0687$$

$$PA = [0.032, 0, 0.285, 0.113, 0.207, 0.043, 0.055, 0.081, 0.141, 0.222, 0, 0]$$

$$U = [0.875, 0, 0.008, 0.022, 0.161, 0.094, 0.064, 0.104, 0.415, 0.014, NaN, NaN]$$

$$\hat{\kappa} = -0.00096$$

Από τα τρία παραπάνω μοντέλα, ελαφρώς καλύτερο είναι το μοντέλο για $\boldsymbol{NR = 2}$, το οποίο βγάζει και αυτό απογοητευτικά αποτελέσματα.

Δεύτερο Μέρος

Το δεύτερο μέρος της εργασίας πραγματεύεται την εφαρμογή μιας πιο συστηματική προσέγγιση στο πρόβλημα της χρήσης ασαφών νευρωνικών μοντέλων σε προβλήματα ταξινόμησης, 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. Επιλέγεται το isolet dataset.

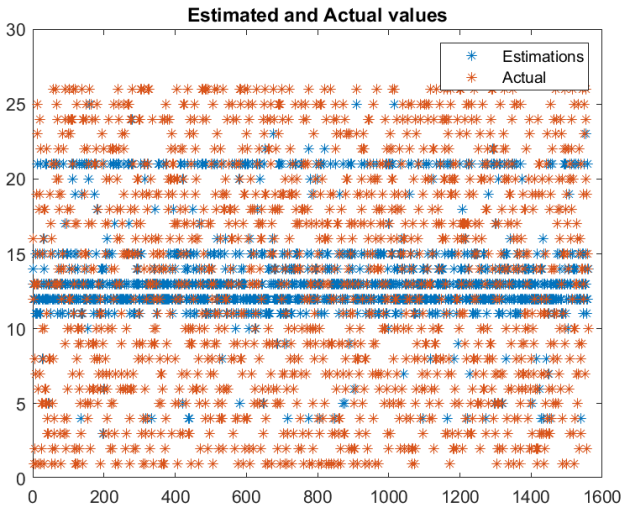
Αρχικά, εφαρμόζεται ο αλγόριθμος Relief ώστε να τοποθετηθούν τα features των δειγμάτων στη βέλτιστη σειρά. Σε κάθε επανάληψη της πεντάπτυχους διασταυρωμένης επικύρωσης γίνεται τυχαία αναδιάταξη των δεδομένων και διαχωρισμός τους και training set (80% των δεδομένων) και test set (20% των δεδομένων) με τη βοήθεια της συνάρτησης `cvpartition()` του Matlab. Για τη δημιουργία του μοντέλου χρησιμοποιείται η συνάρτηση `genfis2()` του Matlab, η οποία χρησιμοποιεί τον αλγόριθμο Subtractive Clustering 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.

Να σημειωθεί πως αντιμετώπισα τα ίδια προβλήματα με το dataset του πρώτου μέρους της εργασίας. Συνεπώς, προκειμένου να μπορέσει να εκπαιδευτεί το μοντέλο και να έχω αποτελέσματα, αντί γαι την παράμετρο NR **μετέβαλλα την παράμετρο randii**, η οποία έχει άμεση εξάρτηση με το πλήθος των κανόνων.

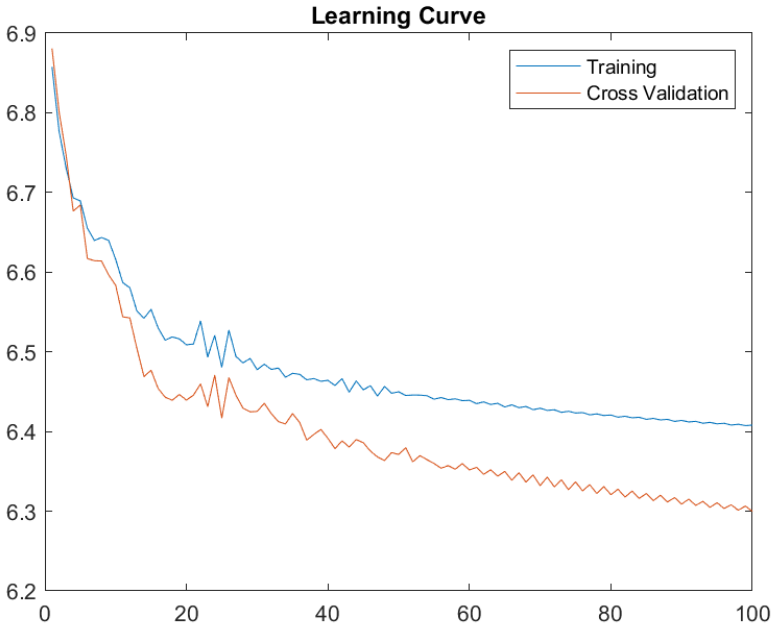
NF / **randii**	3	6	9	12
0.18	7.079	6.9409	6.9631	NaN
0.22	7.052	6.6043	6.9715	7.0219
0.26	7.0904	6.6374	6.5238	6.9713
0.3	7.2175	6.4868	6.5152	6.5376
0.34	7.0707	6.4582	6.4843	6.5030

Πίνακας 4: Μέσο σφάλμα πεντάπτυχους διασταυρωμένης επικύρωσης

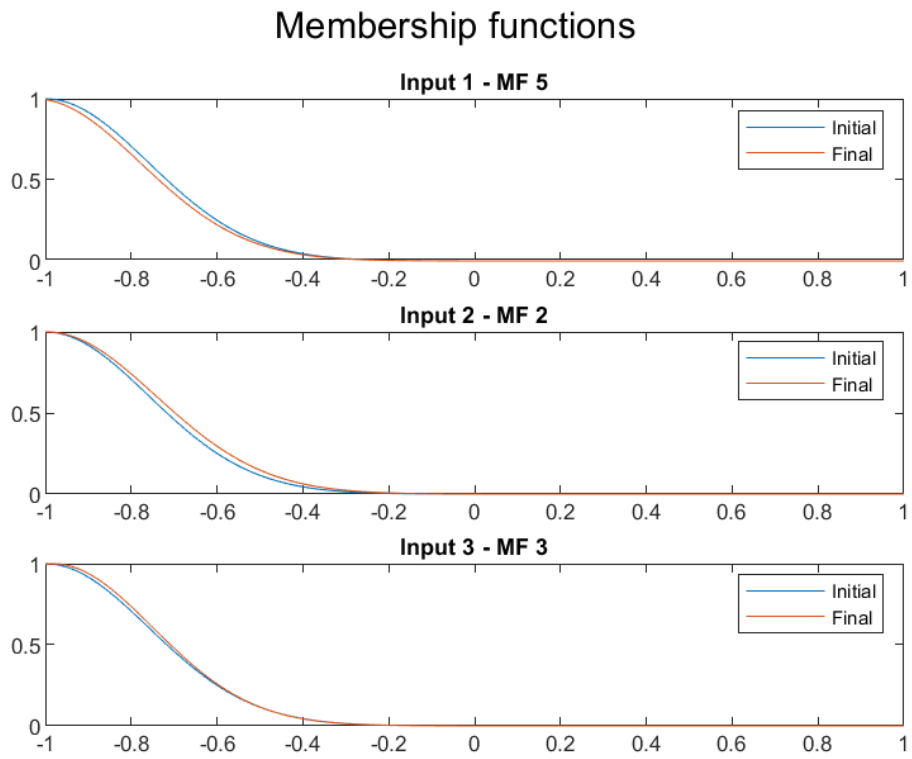
Επιλέγονται, λοιπόν, ως βέλτιστες επιλογές οι $\boldsymbol{NF = 6}$ και $\boldsymbol{randii = 0.34}$. Στη συνέχεια, γίνεται ξανά εκπαίδευση του μοντέλου με τα βέλτιστα χαρακτηριστικά. Τα διαγράμματα που προκύπτουν παρουσιάζονται στα σχήματα 7 έως 9.



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. Επιπλέον, να σημειωθεί πως τα αποτελέσματα είναι και πάλι απογοητευτικά. Τέλος, οι ζητούμενες μετρικές σφάλματος που προκύπτουν από την παραπάνω διαδιασία είναι:

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	26	0	0	0	1	0	0	0
0	0	19	0	0	0	0	0	0	0
0	0	6	0	0	0	2	0	0	0
0	1	2	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	1
0	2	1	1	0	0	0	0	0	2
0	3	1	1	0	0	0	0	0	0
0	2	0	1	0	0	0	0	0	2
4	7	1	11	8	3	0	6	5	0
4	32	0	30	32	39	0	36	33	0
3	13	0	1	13	21	22	17	21	21
2	0	0	0	1	1	18	4	2	21
0	0	1	1	3	2	20	0	4	13
0	0	1	0	0	1	0	0	0	0
0	0	1	1	1	0	0	0	0	0
0	0	0	0	1	1	0	1	1	0
1	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0
2	3	0	0	1	1	2	0	2	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Πίνακας 5: Πίνακας σφαλμάτων

$$OA = 0.0647$$

UA_1 =
NaN
NaN
0.9630
0
0
0
0
0
0
0.2500
0.0089
0.0387
0.0405
0.0250
0.0072
0.2308
0.1000
0.1429
0
0
0.1436
0
0
0
0
1.0000

PA_1 =
0
0
0.4333
0
0
0
0
0
0
0.0333
0.0233
0.3462
0.2424
0.0469
0.0179
0.0536
0.0137
0.0156
0
0
0.5532
0
0
0
0
0.0500

Αν είχαμε επιλέξει grid partition με δύο ή τρία ασαφή σύνολα στην είσοδο θα είχαμε 2^{618} και 3^{618} κανόνες αντίστοιχα. Η πολυπλοκότητα του μοντέλου θα ήταν τελείως ασύμορφη και η υλοποίηση πρακτικά αδύνατη!